Doc.
C 3.238/2:
6

SMALL-AREA STATISTICS PAPERS

SERIES GE-41, No. 6

AN EVALUATION OF SMALL-AREA DATA FORECASTING MODELS AND 1980 CENSUS— SMALL-AREA STATISTICS PROGRAM

Papers Presented at the Conference on Small-Area Statistics

American Statistical Association
Washington, D.C.
August 14, 1979

U.S. Department of Commerce
BUREAU OF THE CENSUS

SMALL-AREA STATISTICS PAPERS

SERIES GE-41, No. 6

AN EVALUATION OF SMALL-AREA DATA FORECASTING MODELS AND 1980 CENSUS— SMALL-AREA STATISTICS PROGRAM

Papers Presented at the Conference on Small-Area Statistics

American Statistical Association
Washington, D.C.
August 14, 1979

Issued May 1980

U.S. Department of Commerce
Philip M. Klutznick, Secretary
Courtenay M. Slater, Chief Economist

BUREAU OF THE CENSUS
Vincent P. Barabba, Director

BUREAU OF THE CENSUS

Vincent P. Barabba, Director

Daniel B. Levine, Deputy Director
Rex L. Pullin, Associate Director
for Field Operations

GEOGRAPHY DIVISION
Jacob Silver, Chief

Library of Congress Cataloging in Publication Data

Conference on Small-Area Statistics, Washington, D.C., 1979.
An evaluation of small-area data forecasting models and 1980 census—small-area statistics program.

(Small-area statistics papers : Series GE-41 ; no. 6)
"Issued in May 1980."
Supt. of Docs. no.: C 3.238/2:6
1. United States—Statistical services—Congresses. 2. United States—Census, 20th, 1980—Congresses. I. American Statistical Association. I. Title. III. Title: Small-area statistics program. IV. Series.
HA37.U55C64 1979 353.0081'9 80-607162

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Postage stamps not acceptable; currency submitted at sender's risk. Remittances from foreign countries must be by international money order or by draft on a U.S. bank.

PREFACE

This report contains the papers presented at the Committee on Small-Area Statistics in Washington, D.C., on August 14, 1979, during two sessions of the annual meeting of the American Statistical Association (ASA), which was held jointly with the Biometric Society, and the Institute of Mathematical Statistics.

The first session of the 1979 meetings concerned *An Evaluation Of Small-Area Data Forecasting Models.* John H. Morawetz organized and chaired this session. The speakers were R. William Thomas, Frank E. Hopkins, and Boyd L. Fjeldsted. Joseph W. Duncan served as discussant.

The second session dealt with the *1980 Census—Small-Area Statistics Program.* Irving Roshwalb organized and chaired this session. The speakers were Marshall L. Turner and Michael G. Garland. Pearl M. Kamer and Edward J. Spar served as discussants.

This report was organized and prepared under the direction of Jacob Silver, Chief, Geography Division, Bureau of the Census. Assisting in the preparation was Nancy James.

FOREWORD

The papers included in this report are the result of a joint effort of the Committee on Small-Area Statistics of the American Statistical Association and the Bureau of the Census. The Committee on Small-Area Statistics was established in 1931 (then called the ASA Committee on Census Enumeration Areas). Very early in its history the Committee established an annual conference held as part of the annual convention of the ASA. The importance of the papers presented at these conferences induced the Bureau of the Census to publish them in a separate publication with a wider circulation than the ASA proceedings. The Bureau has been publishing the papers since 1958.

For its purposes the Committee on Small-Area Statistics defines any area smaller than a State as a small area. In particular the interests of the Committee include:

- Methods of data creation; i.e., methods used to generate small-area statistics by public and private producers of data.
- Methods used in updating decennial census data; evaluation of methods used in intercensal small area estimating by government agencies, private census data suppliers, and users.
- Private and public need for data; how is the need for data determined by the Bureau of the Census, other government agencies, and private suppliers?
- The relationship between data needs and data costs.
- Maintaining channels of communication between the producers and users of small-area statistics.
- Provide information on the sources of small-area statistics, both private and public.

The sessions organized by the Committee as a part of the ASA annual conference reflect these interests. Recently methodology and business needs and uses of small-area statistics have been stressed. The Committee pioneered sessions directed to business concerns in the field of small-area statistics.

Every researcher or decisionmaker who needs small area data is aware that comprehensive data for geographic divisions smaller than the State or SMSA are few and far between. Very little guidance is available to the user on the methodology appropriate to a particular problem. When data are available the definitions and methods of data collection are not always known to the practitioner. As a rule government agencies are careful in defining terms and providing detailed information on methods used and the assumptions made in obtaining estimates, but not all data suppliers are so dedicated to precision. Intercensal estimates of various data are made available by some organizations via simple extrapolation of the latest census figures and users of the data have nothing better than "synthetic estimates;" the assumptions used in obtaining those estimates are usually not provided with the data.

Through its activities the Committee on Small-Area Statistics attemps to ameliorate both problems. Through contacts with data suppliers it helps to increase the availability of data for small areas; the sessions at the annual conference and this report guides users to sources of data, provides exposure to the methods used in data collection and the state of the art in methods of using the data to solve specific problems. The two sessions organized by the Committee for the 1979 conference deals with the availability of data from the 1980 Census (session chaired by Irving Roshwalb) and methods used in small-area forecasting models (session chaired by John Morawetz).

The session on the 1980 Census dealt with the Bureau's plans for that census, and the services the Bureau provides to users. It also provided a forum for users to voice their special problems.

At the other session three papers on forecasting small area activities were presented. One dealt with theoretical matters and the other two illustrated practical model construction problems. Particularly instructive is the Hopkins paper which examines the many pitfalls awaiting the inexperienced researcher.

I hope that readers will find the papers and the accompanying discussion interesting and instructive. If you have any comments on these papers, or suggestions to the Committee, please forward them to me.

Jonah Otelsberg, Ph.D., Chairperson,
Committee on Small-Area Statistics
American Statistical Association

Digitized by the Internet Archive
in 2013

http://archive.org/details/smallareastatist6197unit

CONTENTS

AN EVALUATION OF SMALL-AREA DATA FORECASTING MODELS

1980 CENSUS—SMALL-AREA STATISTICS PROGRAM

An Evaluation of Small-Area Data Forecasting Models

Introduction

John H. Morawetz
McGraw-Hill Information Systems Company

The forecasting of economic events is an absolute necessity for governments, business enterprises, etc. They have a responsibility to plan ahead, prepare budgets, etc. It always requires assumptions about war or peace, changes in human behavior, or the state of the arts, perhaps 25 years hence.

Should a government agency authorize the construction of an office building in a given location? Or, should a utility company build a nuclear power plant, knowing that up to 15 years will go by before it will produce energy? What do they need to know? These are typical questions to which decision-makers need answers.

If professionals have one overriding responsibility—other than training the next generation of professionals—it is to help individuals and organizations to make wise decisions. Therefore, statisticians have come armed with an ever increasing number of data bases, ever improving econometric techniques, and faster computers. In their eagerness to help, they have compromised the "law of large numbers" and have made heroic assumptions about the applicability of national phenomena to small areas. The lengths of the forecasts have likewise been extended.

In recent years I have seen studies that suggest the predictability of selected economic events in areas as small as SMSA's—up to 25 years into the future. Those studies notoriously lack adequate estimates of the irregular components of the time series involved, nor do they include standard error forecasts.

As much as we want to encourage the development of new and better statistical methods, we have an obligation not to profess knowledge we don't possess.

Multitudes of laws and regulations protect an unaware public against harm from conditions that they are unable to evaluate. Building codes, consumer protection laws, especially those enforced by the Food and Drug Administration, are some examples.

Not being regulated, statisticians have an obligation to police their own activities. The Office of Statistical Policy has that responsibility to which federal funds are involved. There is a definite need for the profession to develop and generally accept statistical standards regarding small area and/or long-term forecasts.

This introduction is not intended to take away from the very thoughtful papers presented here in which the various authors exhibit their application of statistical methodologies, pointing the way to a better effort. Rather, it highlights the needs for users of projected data to have a full understanding of their limitations and for the profession to assure that data which fail to meet minimum standards of acceptability are not generated.

A Model of Construction Activity in Subnational Areas

R. William Thomas and H.O. Stekler
Institute for Defense Analyses
Jack L. Rutner
Joel Popkin and Company

INTRODUCTION

This paper presents preliminary results obtained from current research that the Institute for Defense Analyses is doing for the Department of Labor.[1] We are engaged in the task of developing a computerized information system that will provide analyses of construction labor market conditions in local areas. Construction Labor Demand System (CLDS) will provide estimates of current and future construction activity and the labor requirements associated with those activities for each region. CLDS is designed to provide information for specific types of construction and for precise construction trades. This information can be used for analyses of local labor market conditions in the construction industry, and can also assist in the formulation and implementation of policies which would contribute to the efficient operation of these labor markets.

Estimates for construction activity are generated for 29 different construction types and the labor requirements of each type are generated for 29 crafts. For each construction type, estimates of starts are obtained either as data from F. W. Dodge or as forecasts from the model we are developing. Using information on average duration of construction activities, estimates of value put in place are obtained for each specific construction type. Finally, historically observed patterns of the utilization of various construction crafts are used to forecast the labor requirements associated with the particular type of construction activity. The total labor requirement for each of the 29 crafts is obtained by summing across all 29 types of construction activity.

Formally, CLDS consisted of a block recursive system of equations. The set of equations (1) relates starts of particular types of construction to particular demographic and economic variables. Thus,

$$S_t^i = f^i(X_t) \quad , \tag{1}$$

where S_t^i refers to starts of the i^th^ construction type and the t subscript represents time periods. X is a vector of observations on specific economic and demographic variables, including lagged values of the dependent variables.

Starts are then translated into economic activity by assuming that this activity is distributed over time. Consequently,

$$A_t^i = D^i(L)[S_t^i] \quad , \tag{2}$$

where A_t^i is activity (value put in place) of the i^th^ construction type, and $D^i(L)$ is a polynomial lag operator function.

Finally, employment in each of the specific construction trades is obtained by relating labor input requirements to the level of construction activity. Therefore,

$$E_{jt}^i = A_t^i W_j^i \quad , \tag{3}$$

where E_j^i represents the employment of the j^th^ craft by the i^th^ construction sector, and W_j^i is the input requirement for the j^th^ craft in the i^th^ construction sector.

The data sources for starts and the exogenous variables used in equation (1) will be described below. The lag operators D_j^i (L) and the weights W_j^i were obtained from AMIS Construction and Consulting Services, Bureau of the Census, and Bureau of Labor Statistics.

The Institute for Defense Analysis' (IDA) task was to develop the forecasting model which determines the annual level of starts[2] of various types of construction, (see equation (1)).

CHARACTERISTICS OF THE CONSTRUCTION STARTS MODEL

The complexity of building a construction starts model at such a detailed level required the development of some criteria which the model would have to satisfy. First, the model would have to conform as closely as possible to current economic theory about the determinants of construction investment. Thus, the model would be structural in nature and would be suitable by econometric techniques. Second, it was recognized that IDA had neither the time or resources to develop an entire multiregional model. It was undesirable to build a construction model in which the regional determinants of construction investment were treated as exogenous.[3] These criteria led to the conclusion that the construction starts model would be utilized in conjunction with an existing multiregional forecasting model.

The National Regional Impact Evaluation System (NRIES) was chosen from the few multiregional models which were available. This model was developed by the Regional Economic Analysis Division of the Bureau of Economic Analysis (BEA). It is a set of 51 State models (including the District of Columbia) estimated from time series data using a common theoretical structure. Each State model is composed of 69 behavioral equations[4] and generates the State forecasts which provide the regional inputs to our construction model.

The choice of a regional model also effectively determined that the construction starts model would be estimated using data aggregated at the State level. Consequently, the task was to

[1] "Regional Forecasts of Construction Activity Levels," Contract J-9-E-8-0031, U.S. Department of Labor.

[2] IDA also has the task of distributing these annual starts to the specific months of that year and to specific size classes within each construction type. However, we shall not deal with these topics in this paper.

[3] If these variables were treated as exogenous, assumptions about the future values of these variables would have been required for each region for every forecast run.

[4] See Kenneth P. Ballard and Robert M. Wendling, "The National-Regional Impact Evaluation System: A Spatial Model of U.S. Economic and Demographic Activity." Paper presented at the meeting of the Regional Science Association, Chicago, Ill., November 1978, mimeo.

specify a collection of model(s) which would forecast starts for 29 different types of construction for each of the 51 States.

The data required to build this model came from the NRIES data bank and the F. W. Dodge Company. The NRIES data bank provided most of the regional explanatory variables. The F. W. Dodge data are time series data on individual construction starts. These data were aggregated to the State level to form the dependent variables whose behavior was estimated by the model.

CONCEPTUAL QUESTIONS IN MODELING CONSTRUCTION STARTS

In reviewing the existing literature on the determinants of construction investment, it was discovered that the previous research could provide guidance in specifying only a few of the 29 construction types. There have been many analyses of residential housing starts and fixed private investment at the national level. However, little research has been published about the determinants of the specific construction components of fixed private investment even at the national level. Even less information was available about the determinants of nonbuilding construction such as highways, airports, etc. For all of these construction types even less is known about regional determinants of investment. Consequently, new theoretical specifications had to be developed; however, we tried to formulate the equations in a manner consistent with the findings of previous research on the determinants of national investment in structures.

Several broad themes run through the existing investment literature. First, most studies indicate that there is a desired capital stock, but there is not a general agreement about the determinants of these desired stocks. The desired stock might be proportional to output (perhaps taking into account the prices of the outputs and the price of capital services). Alternatively, the desired stocks might depend upon non-output considerations, e.g., internal and external financial variables. Second, it is now generally agreed that when there is a difference between the desired and actual capital stocks, the adjustment is not instantaneous and the discrepancy is removed over time. These findings give rise to the flexible accelerator model, in which new investment starts are a proportion of the difference between the desired and actual stocks.

$$\begin{aligned} K_t^x &= \alpha X_t \\ I_t &= \lambda(K_t^x - X_{t-1}) \end{aligned} \tag{4}$$

K_t^x and K_t represent the desired and actual stocks, respectively. X_t is a variable which determines the desired capital stock, and I_t is investment.

It is one thing to specify an equation and another to have the data which would permit the actual estimation. However, for most construction categories, there are no capital stock figures available by State. Because of the lack of specific data, specifications based on the flexible accelerator principle could not be used in the majority of cases. In other instances, other desired data were not available but proxy variables could be substituted.

Except for the questions of specification and data, the remaining issue is which econometric procedure should be used to estimate the model. The availability of data limits the types of estimators that can be used.

The Dodge data were only made available for the 6 years 1972-1977. Consequently, there were insufficient observations to justify a pure time series approach. While for each year there were 51 regional observations for each construction type, a pure cross-section approach was also rejected. This approach was precluded because the cyclical variability of the construction industry has been so great that no single year within the 1972-1977 period could be viewed as typical. Thus, to capture both the regional and cyclical variability, the data were pooled, thereby obtaining 306 observations for each construction type.

Three econometric techniques have been commonly used with pooled data. They are ordinary least squares, dummy variable regression, and the variance-components methodology.[5]

In the straight OLS approach no significance is attached to the fact that six observations come from each State, and all observations are treated alike. The dummy variable regression is expressed as—

$$Y_{it} = a_i + X_{it}\beta + u_{it} \qquad \begin{array}{l} i=1,\ldots,N \\ t=1,\ldots,T \end{array} \tag{5}$$

where N is the number of years and T is the number of periods. It is assumed that the slope coefficient (β) is the same for all regions and that only the constant terms (a_i's) differ. This implies that each region has certain (unexplained) unique characteristics. In the newer variance-components model, the form of the equation to be estimated is the same as in the dummy variables regression, but the assumptions are different. It is now assumed that the intercepts are random coefficients rather than fixed parameters.[6] These procedures will eventually be applied to all 29 equations, but for now we shall only present the results that were obtained for the single family residential housing starts equations.

THEORETICAL MODEL SPECIFICATION

Theoretical Basis

Our theoretical specification of the single family housing equation is based on the model developed by Muth.[7] Muth posits that the demand for housing stock is proportional to the demand for housing services, which depends on permanent income, relative prices, and the real interest rate. From this he derives a demand function for new housing investment,

[5] The three techniques and their limitations are discussed in detail by G.S. Maddala, *Econometrics*, McGraw-Hill, New York, 1977, pp. 320-330.

[6] Thus, the residual variation around the predicted values of Y may be partitioned into the variation due to differences in α_i ($T\sigma^2_\alpha$) which we refer to as the between-state variation, and the variation due to u_{it}(σ^2_u), the within-State variation.

[7] R.F. Muth, "The Demand for Nonfarm Housing," in Arnold E. Harberger, ed., "Demand for Durable Goods," University of Chicago Press, Chicago, 1960.

which is a function of the determinants of the demand for housing services, the existing stock of housing, and the speed of adjustment of stock to demand charges.

Our empirical specification draws on the work of Maisel[8] as well as the equations used in the major macroeconomic models. Following Jaffee and Rosen,[9] however, we give greater emphasis to demographic variables in our State-oriented model than do those who are estimating national equations from time series data. The basic model specification discussed above may be written as—

$$\begin{aligned} SFS_{it} &= \alpha_i + \beta_1 PMN_{it} + \beta_2 P65_{it} \\ &+ \beta_3 INT_t \cdot P_{it} + \beta_4 DPI_{it} \qquad (6) \\ &+ \beta_5 KSF_{i,t} + \varepsilon_{it} \end{aligned}$$

where

SFS = single family housing starts (units)
PMN = net immigration of population
P65 = population 65 years of age or older
INT = long term interest rate
P = total population
DPI = disposable personal income
KSF = stock of single family units (beginning of period)
ϵ = the error term.

The i subscript represents the State, while the t subscript runs over the years 1972-1977.

This specification follows the conventional stock adjustment model. We were fortunate that measures of housing inventory were available for 1970 and 1976; from these, annual estimates of the housing stock were constructed for the intervening years by the perpetual inventory method. Coefficient β_5 may be interpreted as the speed of adjustment or percent completed per year.

The remaining variables are conventional determinants of the desired stock of housing. Population growth (PMN) adds directly to the number of households. The elderly (P65) tend to prefer multiunit housing; thus, we would expect β_2 to be negative. The interest rate increases the cost of homeownership; thus, β_3 should be negative.[10] Disposable income (DPI) increases will increase the demand for single family homes relative to multifamily units.

Normalization Considerations

Equation (1) would be expected to be subject to considerable heteroscedasticity if estimated directly. Annual values of starts for the various States range from 165,000 in California to 150 in the District of Columbia. Thus, some normalization rule is appropriate. Population is the natural choice; however, another consideration is involved which led us to a different approach.

Each State's total starts may be viewed as a sample from repeated drawings from a population governed by the model above. The larger the population of the State, the larger the (potential) sample—therefore, the sample mean (or sum) should have lower absolute (relative) variance.

One can incorporate this improvement in accuracy by weighting larger States more heavily in the estimation. The appropriate weighting factor is the square root of total population (P). The net result of deflating all variables by P, then weighting observations by the square root of P, yields the form in which the equation was actually estimated:

$$\begin{aligned} (SFS/\sqrt{P})_{it} &= \alpha_i/\sqrt{P}_{it} + \beta_1 (PMN/\sqrt{P})_{it} \\ &+ \beta_3 INT_t \cdot \beta_4 (DPI/\sqrt{P})_{it} \qquad (7) \\ &+ \beta_5 (KSF/\sqrt{P})_{it} + v_{it} \,. \end{aligned}$$

In the above equation, $v_{it} = (\epsilon_{it}/\sqrt{P_{it}})$ is assumed to be distributed normally with

$$\begin{aligned} E(v_{it}) &= 0 \\ E(v_{rt}, v_{su}) &= \sigma_v^2 \text{ if } r=s \text{ and } t=u \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Thus, errors are assumed temporally uncorrelated, spatially uncorrelated between States, and homoscedastic.

ESTIMATION RESULTS

As noted above, three estimation procedures are possible. We denote these as:

1. Simple least squares (OLS),
2. Least squares with individual intercepts (LSDV)
3. Variance-components (VC).

In addition, two choices for pooling were examined. In the first case, data for all 50 States and the District of Columbia were pooled (Case N). The second procedure (Case R) was to separate the data according to census regions and estimate separate coefficients for each region. This procedure was motivated by the observation that the same model might not be appropriate for high growth regions (the Sunbelt) and regions experiencing net outmigration (the North). In particular, the potential for asymmetric behavior with respect to the adjustment of desired actual stock seems likely to be realized for long lived assets such as housing.

For each of the two cases, six sets of estimates could be derived—OLS, LSDV, and VC. Our approach will be to compare the estimators when data are pooled nationally, then to compare regional and national estimates to examine the question of tests for the appropriateness of pooling data.

[8] Sherwin J. Maisel, "A Theory of Fluctuations in Residential Construction Starts," *American Economic Review*, Vol. 53, (June 1963), pp. 359-383.

[9] Dwight Jaffee and Howard S. Rosen, "A Long Run Model of Household Formation, Housing Starts, and Mortgage Financing," Princeton University Press, 1978, mimeo.

[10] Since the other variables are affected by scale, INT has been multiplied by total population (P) to avoid incommensurate values.

Estimates Using Nationally Pooled Data

The first two columns of table 1 present the OLS and LSDV estimates of the parameters of equation (6). Comparison of the standard errors of the two runs indicates that each State variation accounts for a considerable fraction of the residual variation of the OLS run. An F-test for significant variability in the regional intercept terms was performed. The resultant statistic was 21.3. With 50 degrees and 199 degrees of freedom, the critical value of F at the 5-percent level is less than 1.5.

The impact of adding individual intercepts for each State on the coefficient's values is quite marked. The magnitudes of the income, elderly population, and stock variables increase in the LSDV version, while the interest rate and population migration coefficients decline. The sign of the stock coefficient is theoretically incorrect in both versions. All other variables have their theoretically predicted signs.

Using the residual sums of squares from the OLS and LSDV equation, we can estimate θ, the weighting factor associated with the variance-components model. This calculation is illustrated in table 2. θ is the ratio of within-State residual variation to total residual variation.[11] The first column of table 2 contains the residual sum of squares from the OLS equation, which includes both the between- and within-State variation. The second column is the within-State residual sum of squares from the LSDV equation. From this we estimate σ_v^2 and σ_a^2 by—

$$\sigma_v^2 = \frac{\text{RSS-(LSDV)}}{\text{TN-K-N}} \quad ;$$

$$\sigma_a^2 = \frac{\text{RSS-(OLS) - RSS-(LSDV)}}{\text{T(N-1)}}$$

We note that, for the nation, between-State variation dominates the total variation, resulting in a theta of 0.045.

Variance-Components Estimates—National Run

We first note that the VC estimates for the nationally pooled sample are distinguishable from both the OLS and LSDV results, this finding is remarkable. In many applications, it has been found that VC estimates tend to approximate LSDV estimates, as the spatial variation component totally dominates the temporal effect.[12] Evidently, our housing investment data are sufficiently variable over time to prevent this.

The VC coefficients tend to indicate a larger response to interest rate and migration changes, and a lower stock coefficient, than the corresponding coefficients of the OLS and LSDV equation. However, the coefficient of the elderly population is positive in the VC estimates, casting some doubt on the reliability of the specification.

Regional Estimates

Table 3 presents coefficient estimates for each of the four major census regions. Table 2 reports the calculation of θ for the variance-components model estimates. First, the Northeast region had the lowest σ_a^2 (3.61) and was the only region where the variance components were approximately equal. While generalizations after the fact are always dubious, we take this to be some evidence that the States of the Northeastern region are more homogenous with respect to housing investment than those of other regions. The Northeastern States are the most geographically concentrated, the most densely populated, and the slowest growing. Residual variation is quite low for all three estimators in the Northeastern States.

Table 1. Estimates of the Single-Family Home Equation Using Data Pooled Over the Nation

Independent variable	OLS	LSDV	VC
Long term interest rate. . . .	-0.06347 (-5.34)	-0.01731 (-1.85)	-0.1126 (13.2)
Disposable personal income.	0.08334 (4.65)	0.3892 (7.00)	0.1280 (5.33)
Net immigration of population	8.970 (9.24)	3.284 (4.40)	12.40 (7.80)
Population 65 years of age or older	-1.376 (-2.62)	-4.873 (-3.43)	2.224 (2.31)
Stock of single family units .	0.03120 (13.2)	0.1328 (5.38)	0.02857 (5.66)
Construction	5.202 (3.86)	[a]-3.64 [b](21.3)	12.24
R^2	0.714	0.944	0.903
Standard error	9.46	4.19	5.47

[a] Mean of estimated regional effects.

[b] F-statistic associated with composite hypothesis $\alpha_i = \bar{\alpha}$, all i.

We may summarize the empirical results of the regional estimates briefly. The interest rate coefficient is consistently negative and significant in all regions and for all methods. Most other variables' coefficients tend to change signs or become insignificant for certain methods and regions. In particular, we note that stock coefficients generally tend to be positive, in contradiction to theory but consistent with the national results reported above.

Standard errors of estimate are relatively low in the Northeastern and North Central regions, when compared to the Sunbelt regions of the South and West. However, R^2's are comparable for all regions, and given the cross-sectional nature of the data and the inherent variability of housing starts it is quite high.

[11] The appendix on page 8 describes the calculation of the variance components estimator.

[12] Maddala, *op. cit.*, p. 328.

Table 2. Calculation of Theta

Region	RSS-(OLS)	RSS-(LSDV)	N	$T\sigma^2_\alpha$	σ^2_v	σ^2_α	θ
National . . .	22275.3	35000.1	51	375.5	17.64	75.10	0.045
Northeast. . .	247.3	102.9	9	19.0	3.39	3.61	0.155
North Central	1274.6	183.5	12	99.1	4.26	19.80	0.040
South	5837.6	731.9	17	319.1	11.62	63.80	0.035
West	7561.6	1469.1	13	507.7	31.26	101.50	0.058

Test for Pooling

Given the variation of coefficients across regions, it is reasonable to question whether data should be pooled nationally. Comparisons of national and regional residual variation, using an F-test, rejected the hypothesis that the regions share a model with common intercept. The value of the F-statistic for the OLS equation was 6.32, with 18 degrees and 231 degrees of freedom. For the LSDV equation, the result was an F-statistic of 4.91, with 15 degrees and 184 degrees of freedom. Both are significant at the 99 percent confidence level.

SUMMARY AND CONCLUSIONS

We have examined the problem of estimating the investment function for single-family homes from limited time series data for individual States. Three estimation procedures were applied to data pooled both nationally and regionally. Thus, evidence is available on the effects of choosing alternative estimation procedures and on the effects of grouping data differently.

The general model used variance components logic in specifying the nature of the disturbances. Magnitudes, signs, and significance of coefficients proved very sensitive to the choice of estimation method in our study. Some of this variation is due to misspecification of the underlying model. However, researchers should be aware that their conclusions may be an artifact of their estimation method.

In particular, certain of our independent variables (such as the interest rate) varied little across States, but exhibited high intertemporal variance in our sample period. Other variables (such as population over 65 years of age) varied little over time but had high interstate variability. Consequently, our OLS and LSDV results were quite different, as was the VC combination of both components.

The choice of estimator is fundamentally rooted in the assumptions of the model and the purpose of the research. If we were interested in the longrun adjustment of housing to demographic and income change, we would alter the model to predict housing stock and use OLS results. Since our concern was with forecasting short-term adjustment dynamics in the housing market, the LSDV model with its assumption of fixed region-specific effects, seems appropriate. If the major concern is to estimate parameters accurately for purposes of hypothesis testing, then the variance-components model is the efficient estimator under the assumption that state intercepts are realizations of a random coefficient. Thus, the choice of estimator should be made on *a priori* grounds and not on the basis of specific estimation results.

Our tests of pooling rejected the hypothesis that the same model applied in all four regions. Since regional results differ so much, the researcher is in a quandry—should he simply record the different estimates and stop, or should experimentation continue to examine the possibility of using different specifications in each region. If the latter course is taken, the complexity of the modeling task is increased considerably. On the other hand, if a common specification is adopted, the likelihood of all coefficients conforming to theory is low. We have no simple answer to this problem.

Applying good statistical methods to small area data is not easy. The data are often crude, and it is easy to be overwhelmed by the data management task. We are encouraged that this exercise showed that sophisticated methods do result in different findings.

Table 3. Estimates of the Single-Family Home Equation Using State Data Pooled, by Region

Independent variable	OLS	LSDV	VC	OLS	LSDV	VC
	The Northeastern States			The North Central States		
Long term interest rate. . . .	-0.06178 (-6.83)	-0.03472 (-1.89)	-0.03125 (-4.19)	-0.05725 (-4.68)	-0.1016 (-10.7)	-0.05091 (-15.8)
Disposable personal income.	0.06762 (5.16)	0.2120 (1.56)	0.1480 (1.64)	0.06847 (3.31)	0.009671 (-0.258)	0.4607 (25.4)
Net immigration of population	3.294 (2.64)	0.1336 (0.117)	-3.629 (-1.64)	4.518 (1.98)	2.233 (2.00)	15.15 (16.1)
Population 65 years of age or older	2.499 (3.22)	5.719 (1.04)	0.5985 (1.26)	1.460 (1.50)	20.58 (2.24)	-4.718 (8.61)
Stock of single family units .	0.0062 (4.29)	0.0964 (-2.53)	0.01217 (3.87)	0.01870 (4.82)	0.02298 (0.451)	-0.02267 (-9.22)
Construction	8.154 (9.35)	[a]1.120 [b](5.44)	10.20	2.622 (1.13)	[a]-1.520 [b](23.2)	7.189
R^2	0.853	0.939	0.836	0.799	0.971	0.990
Standard error	2.518	1.822	2.628	4.858	2.066	1.0325
	The South			The West		
Long term interest rate. . . .	-0.1322 (-6.12)	-0.1236 (-9.05)	NOT CALCULATED	-0.09778 (-2.55)	-0.1409 (-5.80)	-0.09996 (-3.64)
Disposable personal income.	0.1543 (4.85)	0.6398 (6.57)		0.08201 (1.36)	0.4989 (2.97)	0.05787 (1.27)
Net immigration of population	1.034 (0.610)	0.007445 (0.00550)		9.910 (2.64)	6.831 (2.32)	23.52 (4.70)
Population 65 years of age or older	0.009071 (0.0130)	1.711 (0.365)		0.8134 (0.183)	11.29 (1.09)	7.657 (2.85)
Stock of single family units .	0.0481 (9.46)	-0.02209 (0.902)		0.03465 (1.60)	0.1203 (1.43)	0.01087 (0.95)
Construction	-6.053 (1.87)	[a]-0.320 [b](27.5)		8.140 (2.23)	[a]-3.990 [b](16.2)	4.927
R^2	0.812	0.976		0.725	0.947	0.856
Standard error	8.596	3.409		11.32	5.591	7.878

[a] Mean of estimated regional effects.

[b] F-statistic associated with composite hypothesis $\alpha_i = \bar{\alpha}$, all i.

Appendix A. Statistical Methodology for Estimating Pooled Cross-Section and Time-Series Data

Assume that we wish to estimate an equation of the form

$$Y_{it} = a_i + \beta X_{it} + \mu_{it} \qquad \begin{array}{l} i=1,2,\ldots,N \\ t=1,2,\ldots,T \ . \end{array}$$

Define T_{xx}, T_{xy}, and T_{yy} as the total sums of squares and sums of products and W_{xx}, W_{xy}, and W_{yy} as the within-groups of squares and sums of products.

Then

$$\begin{aligned} T_{xx} &= \sum_{it}\sum (X_{it} - \bar{X})^2 \\ T_{xy} &= \sum_{it}\sum (X_{it} - \bar{X})\,(Y_{it} - \bar{Y}) \\ T_{yy} &= \sum_{it}\sum (X_{it} - \bar{Y})^2 \ . \end{aligned}$$

$\bar{X}$ and $\bar{Y}$ are the means based on all observations.

Also,

$$\begin{aligned} W_{xx} &= \sum_{it} X_{it}(X_{it} - \bar{X}_i) \\ W_{xy} &= \sum_{it} X_{it}(Y_{it} - \bar{Y}_i) \\ W_{yy} &= \sum_{it} Y_{it}(Y_{it} - \bar{Y}_i) \ . \end{aligned}$$

Here the $\bar{X}_i$ and $\bar{Y}_i$ are the means within the i^{th} groups.

Then the least square estimator for this *dummy regression technique,* which assumes a common slope but different intercept is

$$\hat{B} = \frac{W_{xy}}{W_{xx}} \qquad \text{and} \qquad a_i = \bar{Y}_i - \hat{B}\,\bar{X}_i \ .$$

The variance components model assumes the same model:

$$Y_{it} = a_i + \beta\, X_{it} + \mu_{it} \qquad \begin{array}{l} i=1,2,\ldots,N \\ t=1,2,\ldots,N \end{array}$$

and make the following further assumptions:

$$\begin{array}{rcll} E(a_i) & = & 0 & E(\mu_{it}) = 0 \\ \mathrm{Cov}(a_i a_j) & = & \sigma_a^2 & \text{for } i=j \\ & = & 0 & \text{otherwise,} \\ \mathrm{Cov}(\mu_{it},\mu_{js}) & = & \sigma^2 & \text{for } i=j,\ t=s \\ & = & 0 & \text{otherwise,} \\ \mathrm{Cov}(a_i,\mu_{jt}) & = & 0 & \text{for all } i,j,t \ . \end{array}$$

Now $V_{it} = a_i + \mu_{it}$ and

$$\begin{array}{rcll} \mathrm{Cov}(V_{it},V_{iS}) & = & \sigma_u^2 + \sigma_a^2 & \text{for } t=S \\ & = & \sigma_a^2 & \text{for } t \neq S \ . \end{array}$$

It can then be shown that the GLS estimator is

$$\hat{\beta}_{GLS} = \frac{X_{xy} + \theta B_{xy}}{W_{xx} + \theta B_{xx}} \qquad \theta = \frac{\sigma_u^2}{\sigma_\mu^2 + T\sigma_a^2} \ .$$

B_{xx}, B_{xy}, and B_{yy} refer to the between group squares and sums of products, i.e.,

$$\begin{aligned} B_{xx} &= T_{xx} - W_{xx} \\ B_{xy} &= T_{xy} - W_{xy} \\ B_{yy} &= T_{yy} - W_{yy} \ . \end{aligned}$$

It is also known that

$$\hat{\beta}_{OLS} = \frac{T_{xy}}{T_{xx}} = \frac{W_{xy} + B_{xy}}{W_{xx} + B_{xx}}$$

and was shown above that

$$\hat{\beta} = \frac{W_{xy}}{W_{xx}} \ .$$

Consequently, the other two estimators are special cases of the GLS estimator with $\theta = 1$ for OLS and $\theta = 0$ for the least square dummy variable technique.

Developing and Managing a Small-Area Forecasting Model—READ

Frank E. Hopkins
Department of Energy

INTRODUCTION

The construction of the Regional Energy Activity and Demographic (READ) model utilized resources from different offices within the Federal Energy Administration (FEA) and the Department of Energy (DOE) and has extended over four major phases: planning, development, implementation, and external review. This report discusses the managerial procedures that were used to allocate financial and human resources to efficiently control the construction of the model.

The planning phase began in the fall of 1975 and ended in June 1976 when formal approval was given to proceed with development. The development phase can be further disaggregated into three time phased efforts: preliminary model, full model, and extension to include an endogenous environmental sector. In order to improve managerial control, each time phase of the development effort has been further disaggregated into four task óriented areas: data base, software, equation estimation, and simulation and forecasting. The review phase began in the spring of 1978 and continued through the winter of 1979.

The implementation phase was also initiated at the completion of the planning phase. A major aspect of the planning phase was the determination of the usefulness of the model in DOE and other Federal agencies. Preparation of appropriate model interfaces and report requirements were begun at this time.

The contract and personnel resources devoted to the READ model are presented in tables 1A and 1B. The resources utilized in the planning phase were 1.1 person-years and $155,000 in contract money during fiscal year (FY) 1976. FY 1977 contained 15 months because of a revision in the fiscal year. During this period, 3.1 person-years from Applied Analysis and 0.9 person-years from energy data and $449,600 in contract funds were used on the model. There were 5 person-years from Applied Analysis and 0.6 person-years from energy data and $65,000 in contract expenditures devoted to further development of the model in FY 1978.

The remainder of the paper will describe how these resources were utilized. The scarce resources in the Office of Energy Information and Analysis in the old FEA and the current Energy Information Administration (EIA), are periodically subjected to priority reallocation in order to respond to critical analyses needs such as the National Energy Outlook (NEO) in 1976, the National Energy Plan (NEP) in 1977, the Administrator's Annual Report (AAR) in 1977, the National Energy Strategic Study (NESS) parts I and II, and the EIA response to a letter from Senator Jackson on provisions of the National Energy Act as amended in 1978. The management philosophy underlying READ resource allocation has been to assist in these efforts by enhancing non-READ model developments when required and modifying the READ development schedule when necessary. A discussion of management control of the READ effort can only be complete if it includes related analysis efforts that have occurred at FEA and DOE. The planning phase is discussed in greater detail as shown below. The resources available since the beginning of the developmental phase will be outlined as shown on page 12. The development and implementation phases are presented on pages 12 and 14, while the review phase is discussed on page 15. The concluding section will discuss the efficiency of the organizational structure of DOE in regard to analysis and model development based upon the historical experience of the READ effort.

The major conclusions of the paper follow:

- The major problems in estimating small area forecasting models are not technical. Deficiencies in software, estimation, and simulation procedures can be corrected when discovered during the development and review phases of the model's construction.
- Data availability provides limitations on the geographical and sectoral detail of a model, but not an absolute constraint on its development. While estimation of a model at the county level may not be feasible at this time because of data limitations, the estimation of a local area model for SMSA's and remainder of States areas appears feasible.
- The major difficulty in developing a large scale small area forecasting model in the Federal government is organizational. The rapid turnover of top management, particularly when a new agency is being formed, results in constant pressure to analyze short turn-around problems and to rejustify earlier model development plans, both of which, reduce resources devoted to model development.

PLANNING PHASE

The planning phase for the READ model began in the fall of 1975 and continued until June 2, 1976, when John Christie, Assistant Administrator of the Office of Energy Information and Analysis (FEA), reviewed a decision memo on regional modeling and gave approval to proceed with model development. The activities in the planning phase can be divided into four areas:

1. Information dissemination on the model to other FEA offices.
2. Model specification.
3. Analysis of data requirements.
4. Resource planning.

Information Dissemination

This activity consisted of distributing written material and holding informal briefings on the proposed model to various

Table 1A. READ Model Resource Utilization

Personnel	Start	End	1976		1977		1978	
			Percent READ	Person months[1]	Percent READ	Person months[1]	Percent READ	Person months[1]
Total				13.5		47.3		66.5
Applied Analysis								
Total				10.2		36.8		59.6
Morlan	10/12/76	[2] 9/28/78			50	6.8	80	9.6
Gamson	4/11/76				80	12.0	80	9.6
McCallister	9/12/76				60	7.5	100	12.0
Rubin	5/26/77	[2] 7/1/78			50	2.0	60	5.4
Tannen (IPA)	8/20/77	[3] 8/30/78			80	1.0	90	9.0
	[4] 8/30/78						100	1.0
Durst	1/29/78						75	6.0
Wagner (summer)	6/20/78	9/1/78					100	3.0
Klemm	7/12/78						33	1.0
Energy Data								
Total				3.3		10.5		6.0
Hong			30	3.3	50	7.5	25	3.0
Disbrow					20	3.0	25	3.0

[1] Months are percent READ multiplied by time in FY spent at FEA or DOE on READ with 15 months in FY 1977.
[2] Transferred to DAD.
[3] IPA ended.
[4] Permanent.

Table 1B. READ Model Contract Funds

Agency	Dollars		
	1976	1977	1978
Energy Information Administration			
Total	155,000	449,600	
Energy Systems Modeling and Forecasting	51,000	164,600	
Conservation and Environment	9,000	75,000	
Energy System Data	95,000	120,000	
Department of Energy			
Total		(NA)	65,000
Applied Analysis		(NA)	33,000
Energy Data		(NA)	32,000

NA Not available.

offices within FEA and the Federal government. A seminar was held on March 11, 1976, in which the READ model and its relationship to the Structural Econometric Energy Demand (SEED) models, which were being scheduled to replace the Midrange Energy Forecasting System (MEFS) reduced form energy demand (RDFOR) models, was presented. Documentation of the READ model was also distributed during the seminar. Responses on the seminar were solicited from representative offices. These responses were used as the basis for modifying the design of the model, so it could effectively interface with existing and planned FEA programs.

Model Specification

The four dimensions that had to be considered in specifying the model were: Time, space, equation coverage, and model interfaces. An interactive process between the staff of the former Demand Analysis Division and other offices in FEA was used to specify the dimensions of the model.

Time—The primary use of the model was to provide inputs into the RDFOR and SEED models for use in the MEFS. Since these models require annual exogenous data, the READ model was designed to generate annual forecasts from the year 1978 to the year 2000.

Space—The four major factors that lead to the decision to use the county as the basic unit to estimate the equations were: Data availability, theoretical structure, forecast flexibility, and forecast accuracy. The data used in estimating the model is primarily economic and demographic. The county is the basic unit for collection of much of the data. Since the demand for energy will not be estimated within the model, county-level energy consumption data is not required. The theoretical structure of the industrial location sector is heavily dependent upon using transportation costs, generated from linear programming transportation problems, as explanatory variables. Theoretically, the transportation costs or shadow prices become less meaningful as a reflection of economic activity, when regional units are aggregated to the State level. The use of counties as observational units provides forecast flexibility by permitting aggregation across multi-county unit areas such as States, DOE regions, intra-State areas, etc. Finally, forecasts at State and regional levels may be more accurate if they are generated from county forecasts.

Equation Coverage—The equation coverage was determined on the basis of data availability and model interfaces. The decision to develop a preliminary model before completing a full scale model was also based upon these considerations. The preliminary model was divided into four interdependent sectors: Industrial, construction, demographic, State, and local government.

Model Interfaces—Since the READ model was designed to interface with a number of FEA/DOE models, the input requirements of these models were a major determinant of the specification of the model. A number of models using READ forecasts as inputs were not scheduled to be completed for a number of years. The 3- and 4-digit Standard Industrial Classification (SIC) process SEED models were in this category. This provided additional justification for restricting the industrial locational equations to the 2-digit level in the preliminary model.

Data Requirements

The process for obtaining data for use in the model acted as a major constraint on model coverage. Since the model is primarily an economic and demographic model not requiring primary source energy data, a decision was made to rely on secondary data collected by other Government agencies and private corporations. The use of secondary data significantly reduced data acquisition cost. In addition, since FEA was a Government agency, limited information or proprietary data was obtained for several series including the BEA employment and income data.

Resource Requirements

A schedule of resource requirements was developed that proposed a staff of 10 and immediate use of contract funds in June 1976 for data acquisition and contract support. Table 1A shows that the maximal staff of 6-person years was not reached until FY 1978, when DOE was organized.

The proposed staff of 10 was divided between the Office of Energy Systems Data which had 4, and the Office of Energy Systems Modeling and Forecasting which had 6. The proposed schedule under this plan was to have the data base development completed by October 1, 1976. The preliminary model was to have been completed in the winter of 1976, while the full model was scheduled for release by July 1977. This schedule was not realized for four primary reasons:

1. Requested personnel and contract funds were not allocated.
2. Personnel recruitment was exceedingly slow.
3. Contract processing was delayed.
4. As a result, the data base for the preliminary model was not completed until the spring of 1978.

RESOURCE AVAILABILITY

The resources that have been utilized in the development phase of the READ model project, as shown in table 1A, differ considerably in level and in time phase from those requested during the planning phase. The planning effort involved relatively few resources and was completed satisfactorily. However, the resources requested for the development phase placed a large burden on existing EIA resources.

The model development was to be divided between the Office of Energy Systems Modeling and Forecasting (ESMF) and the Office of Energy Systems Data (ESD). The request for six new positions to work on READ was not satisfied in FY 1976, and was only fulfilled in November of 1977 with the formation of DOE. Instead of allocating new positions to work solely on READ, the tasks were assigned to the Demand Analysis Division (DAD) staff within ESMF. There was an understanding between DAD and William Donnelly, Director of READ during this time period, that DAD personnel could work on READ as long as it did not interfere with the completion of other activities.

The failure of the DAD organizational system to complete development of the READ model under its original schedule, but also other development efforts is understandable given the general level of uncertainty of the emerging structure of DOE from combining FEA and other agencies. Rational behavior for office directors, division chiefs, and senior staff, who desired to maintain their positions in the new agency, was to perform well on short-term projects requested by the energy office in the White House since there was an absence of long-run FEA planning during this period. This pattern was followed and many long-run modeling efforts, in addition to READ, were given lower priorities.

The old ESMF office was incorporated in the Office of Applied Analysis when DOE was organized. The problem of using the same personnel to develop the model and to perform the regular functions of DAD was potentially corrected when the Office of Applied Analysis (AA) was organized in November 1977. The Office of Energy Use Analysis (EUA) was created, in the Office of Applied Analysis, to assume the functions of the old DAD. Three divisions were created in EUA: Regional Energy Activity, Conservation and Renewable Resources, and Demand Analysis. Thus, the READ model was given a separate structure, which theoretically corrected the conflicting goal resource allocation problem.

DEVELOPMENT PHASE

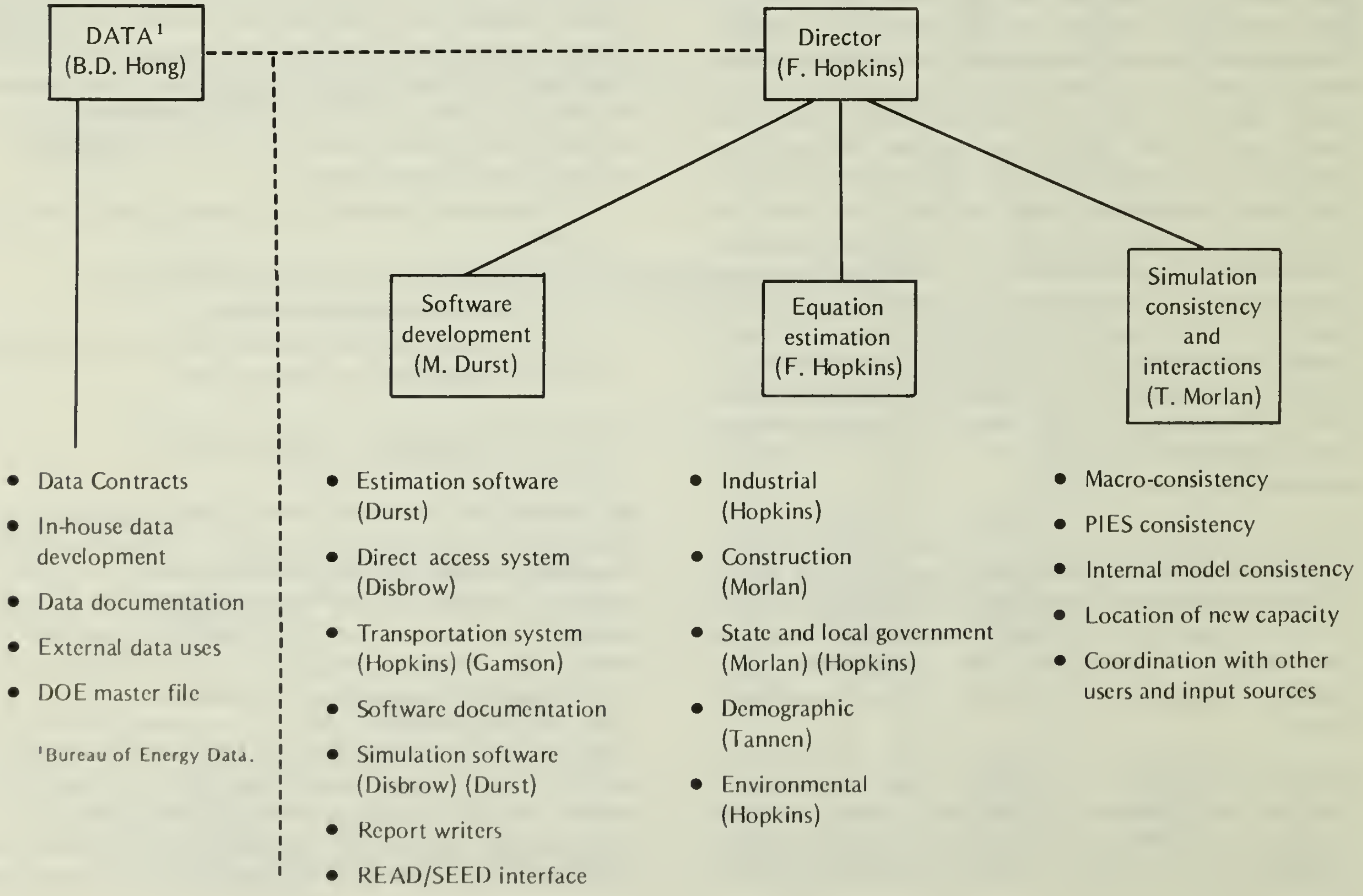
The development of the model has been divided into three major time phased efforts: Preliminary model, full model, and full model extension to include an endogenous environmental sector. Each model phase has been further divided into the four areas described in figure 1: Data, software, equation estimation, and simulation. This section will discuss the first three areas related to the development of the preliminary model.

Data

The goal of the data task was to create a data base that would be:

- accessible to both READ and non-READ staff; and
- definitionally consistent across different data series.

Figure 1. READ Organizational Chart


DATA[1] (B.D. Hong)
Director (F. Hopkins)
Software development (M. Durst)
Equation estimation (F. Hopkins)
Simulation consistency and interactions (T. Morlan)
• Data Contracts
• In-house data development
• Data documentation
• External data uses
• DOE master file
• Estimation software (Durst)
• Direct access system (Disbrow)
• Transportation system (Hopkins) (Gamson)
• Software documentation
• Simulation software (Disbrow) (Durst)
• Report writers
• READ/SEED interface
• Industrial (Hopkins)
• Construction (Morlan)
• State and local government (Morlan) (Hopkins)
• Demographic (Tannen)
• Environmental (Hopkins)
• Macro-consistency
• PIES consistency
• Internal model consistency
• Location of new capacity
• Coordination with other users and input sources

[1] Bureau of Energy Data.

Production of generalized data files extended the time period for completion of the data base over what it would have been if they were created solely for use in READ. While much of these tasks were performed by contracts, certain portions of the data base were processed by DAD staff.

B.D. Hong of the Office of Energy Data maintained the lead role in developing the data for use in both the preliminary and full model of both organizations. Mr. Hong, originally requested a staff of three to assist him in the data development effort. However, resource constraints prohibited the assignment of personnel. The alternative approach was to utilize contract funds and DAD personnel, originally assigned for work on the software and estimated portions of the model, for data development. This approach, while deviating from the original resource request, may have been the most feasible of available alternatives for two reasons. First, widespread support from offices in FEA resulted in the transfer of funds to be used to develop the model. Second, even if personnel were allocated to the data development activity, the time for hiring, training, and completion of the data tasks, could exceed that required by a contractor for delivery of the processed data. Contracts for developing the data base can be divided into two areas: Data acquisition and data processing. The data acquisition contracts involved purchasing the secondary data listed in table 2. The table also lists dates that the data was delivered to FEA. While much of the data were delivered during the summer and fall of 1976, the Dodge Construction reports data were not acquired until May of 1977.

The Orkand Corporation was assigned the data development task of creating the regression file data. Their contract was initiated as of September 1976 and was scheduled to be completed by February 1977. The contract funds were exhausted before their tasks were completed, therefore, the contract was extended for 6 months. Unfortunately, there was a contract procurement delay of 6 months between the end of the initial contract and the start of the extension. Thus, the portion of the regression file data that was to have been processed by the Orkand Corporation was not delivered until February 6, 1978.

Table 2. **READ Source Data Acquisition and Processing**

Data series	Data received	Processing staff
BEA Income Employment[1]	9/76	Orkand
Dodge Construction	5/77	FEA
Export–Imports	7/76	Orkand
Census–Retail Trade Statistics	7/76	FEA
Census–Manufacturing Industries .	7/76	Systems Sciences
Census of Mining	7/76	Systems Sciences
Survey of Manufactures..........	5/76	Orkand
Census of Transportation	6/76	FEA
ICC Revenue Data	5/76	FEA
ICC Cost Data	5/76	FEA
Weather (NOAA)	4/76	FEA
Vital Statistics	12/76	Orkand
Population	1/76	Orkand
Census of Housing	4/76	Synergy
Census of Governments	12/76	Orkand-FEA
County Business Patterns	3/77	Systems Sciences

[1] Originally received January 1976; however, a revised set was not received until September 1976.

The division of labor between the Orkand Corporation, FEA/DOE staff, and Systems Sciences, the successor firm to Orkand, in processing the data is presented in table 2. The Demand Analysis Division staff, occupied with data development, were not able to begin equation estimation during this period. The processing of the regression data file for the preliminary model was not completed until the spring of 1978.

Software

There were four managerial goals in designing the software system: Computational efficiency, data storage efficiency, security, and, generality and ease of use. These goals were achieved, illustrating that the technical problem of managing a large scale small area data base can be solved.

Data storage efficiency of the 40 million data elements of the system were implemented by designing the direct access data file that was used to store data for use in the regression equations. The two major techniques used in the design of the system to achieve minimum storage requirements were: On-line data transformation and data compression A reduction in storage space from 1 disk to 1/3 of a disk was accomplished by compressing the stored data on continuous bit strings, by eliminating missing observations and suppressing zeros on the left of the significant digits of the number.

The specification of a regression equation may consist of three classes of variables—pure, transformed, and hybrid. Only pure variables are stored, since on-line data transformations are used to process the transformed and hybrid variables (see figure 2). By definition, a pure variable does not have to be algebraically transformed before it is used in a regression equation. A transformed variable is a single pure variable that has been modified by an algebraic procedure, such as taking the natural log of a series. A hybrid variable consists of the formation of a single data series from two or more pure or transformed variables. Thus, the value of residential construction is a pure variable, the change or log or residential construction is a transformed variable, while residential construction per capita is a hybrid variable. There would have been approximately 75 million pure variable words requiring disk storage on the full model. The number could easily surpass several billion if all the transformed and hybrid variables were stored on disk. As illustrated in figure 2, the software has been designed so that only pure variables are stored on disk and all transformations on the data are executed as the cross-product matrix is being created and saved for the regression analysis or for use in simulation. A 16 character format for variable identification listed at the bottom of figure 2 has been created to ensure consistency in notation between users of the same independent variables in different equations. The variable identification system also permits definition of an essentially unlimited number of variables.

Estimation

Estimation of the regression equations did not begin until the fall of 1977, when sufficient data and software had been developed to estimate specific sectors of the model. The initial equations for the construction, demographic, State, and local government sector were undergoing testing and further refinement during the summer of 1978. The industrial location sector was the only sector remaining to be estimated by the fall of 1978. The five major managerial goals in the estimation phase were to:

- utilize continued software development to reduce staff time required to estimate the equations,
- create a computer file of estimated equations that could efficiently be utilized in the simulation program,
- effectively utilize specialized staff to estimate a set of equations that could be used for forecasting,
- estimate equations using OLS that would provide an initial set of empirical statements drawn from a previously unexplored set of data, and
- develop full information maximum likelihood techniques for use in the full scale model.

IMPLEMENTATION PHASE

The task of implementing the model began concurrently with the development phase after the planning phase was completed. The implementation phase consisted of four major activities:

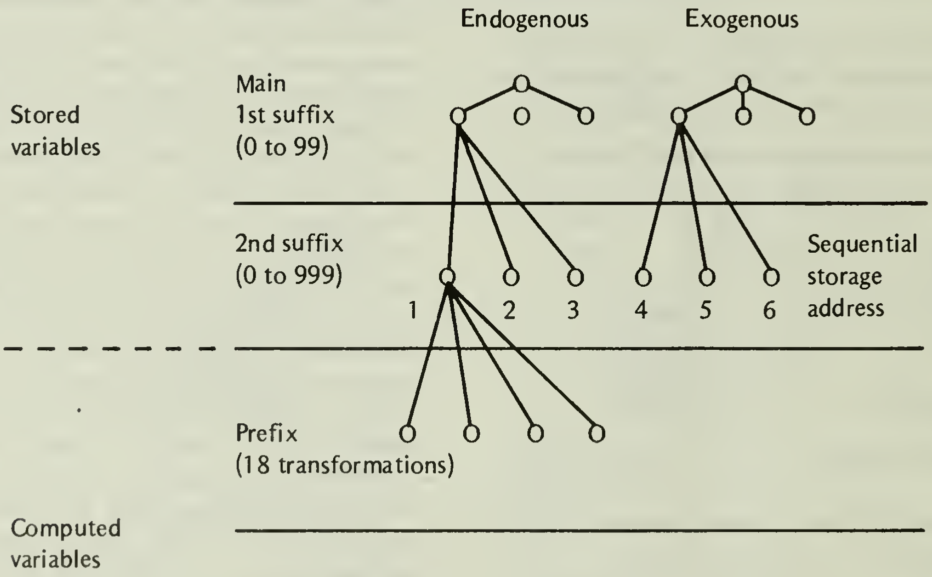
1. Structuring the READ model to interface with other energy impact models (also described in another document "READ Interfaces.")

2. Creation of an interface program between the READ and SEED models (undertaken under contract to the Synergy Corporation that has resulted in the report, "READ and SEED Integrating Systems and Related Studies)."

3. Development of a State energy price forecasting system.

4. Assisting in completing the SEED models, this activity required the services of Frank Hopkins and Lewis Rubin. While this delayed the development of the READ model, the decision was made to reallocate resources to complete the SEED-MEFS interfaces, since they were the primary reason

Figure 2. **Data Storage Configuration In READ**



HYBRID Code Structure for Basic READ Variables:

Item	Characters	Description
Prefix	1 to 2	Transformations
Main	3 to 5	Sector ID
1st suffix	6 to 7	Within Sector Name
2nd suffix	8 to 10	Variable Number
3rd suffix	11 to 12	EX or ED Status
Scale	13 to 16	Scaling Factor for all Pure Variables

for developing READ. The State price forecasting system was initially developed under contract to the Orkand Corporation and modified and upgraded by EUA staff.

Impact Model Implementation

The structure of the READ model was designed and modified during its development to interface the specific models or activities listed below:

- Regional Emissions Projection System (REPS),
- Construction Labor Demand System (CLDS),
- Comprehensive Human Resources Data System (CHRDS),
- Puerto Rican Input-Output Model and,
- Conservation Resource Impact Factors.

SEED Model Development

The organization of DOE, as discussed earlier, resulted in a disruption of model development schedules for a number of projects, including the SEED models. Integrating these models with the MEFS was the responsibility of the DAD. The completion of the 1977 Administrator's Annual Report to Congress (AAR) utilized most of the resources of the DAD until February 1978.

A model development planning document outlining a proposal for future resource allocations within EUA was prepared by Frank Hopkins and submitted to Nicolai Timenes, Jr., Assistant Administrator, Applied Analysis, on March 2, 1978. The general strategy of the model development plan was to concentrate DAD resources on specific tasks that could be completed, while staff was being hired, before work on the 1978 AAR commenced. In addition, since contract funds were available, they were substituted for staff where possible. There were six major model development areas that were discussed in the memorandum: State energy price forecasting system, residential (Hirst) model, commercial model, industrial models, transportation models, and systems development for the AAR. The price forecasting system residential model, and the highway gasoline component of the transportation models were given priority for completion during the spring and summer of 1978. Resources were withdrawn for integrating the commercial model until the fall of 1978, when the Oakridge National Laboratory (ORNL) was scheduled to deliver the DOE regional version of the model. The proposed schedule was followed by shifting personnel from the READ Division and by processing contracts during the spring and summer of 1978.

REVIEW PHASE

The review process can be disaggregated into 11 chronological sequences: Review planning, review contract procurement, READ staff review preparation, first review meeting, reviewers' comments, READ staff response, second review meeting, READ staff response, reviewers' comments, third review meeting, and concluding documentation.

The initial review planning was begun in the winter of 1978 at the request of Dr. Lincoln Moses, Administrator, EIA. The original plan was to utilize the National Science Foundation (NSF) to undertake the study in the summer of 1978. Unfortunately, because the formation of DOE resulted in severe disruptions in the contracting process, the NSF could not be engaged to participate in the review. An alternative review procedure was devised whereby the eight independent consultants listed in appendix A, conducted the review under the organizational direction of George Lady, Director, Office of Analysis Oversight and Access.

READ Staff Preparation

The work by the READ staff in preparation for the review began in June 1978, and consisted of preparing five documents to be sent to the review committee:

- The Regional Energy Activity and Demographic (READ) model: description and applications (46 pages),
- User's guide to the READ regression file (200 pages),
- User's guide to the READ estimation and simulation software (22 pages),
- READ estimation and simulation software technical operating manual (73 pages), and
- READ model validation procedures (51 pages).

The documents, totaling 392 pages, were mailed to the members of the committee on September 27, 1978. Additional documents were being prepared for distribution. These activities occupied over 75 percent of the time of the READ staff from June 1978 to October 1978. The major opportunity cost of these activities was postponement of completion of the model. The equations in all of the sectors, except the industrial, were estimated for completion by the summer of 1978. The remaining equations were scheduled for completion during the fall of 1978. Testing of the simulation software using these equations was also initiated during this period. The simulations were designed to test the robustness of the initial equations, which would be modified when necessary. The preparation for the review precluded the completion of this activity. While the reviewers did provide valuable insight into improving the model structure, the READ staff had speculated that more directly applicable empirical information would have been obtained if the simulation tests had been completed; particularly since the READ staff was aware of the problem areas that the review committee addressed during all three meetings.

First Meeting

The first meeting was held on October 12, 1978. Frank Hopkins made a presentation on the purposes and scope of the READ effort and the review committee members asked questions during the presentation. Their main themes centered on:

1. The county as a meaningful economic unit,
2. The appropriateness of OLS in estimating the preliminary model,
3. The use of synthetic or derived data in the regression equations and,
4. The qualifications and background of the READ staff.

The committee members submitted individual written reports reviewing all aspects of model development after the first meeting. Discussion of these reports was the general topic of the second meeting. Their comments were generally related to judging the scientific integrity of the model against a perfect standard rather than against alternative analytical procedures.

The panel was composed of highly qualified specialists in econometrics, economics, operations research, and statistics. The written comments were general elaborations of the verbal comments discussed in the first meeting. When taken together, the comments appeared to have more serious implications for the model, than if each criticism was analyzed individually.

The READ staff approached the review process from a different perspective than the committee members. While perfection in model development is a desirable goal, a realistic evaluation must consider the model development effort in relation to resource availability and alternative techniques that may be used to achieve the goals of the model. Thus, the READ staff viewed the major issue as a review of the management of resources to achieve the goals described in the Planning Phase section.

A paper describing the management plan of the READ effort and the qualifications of the READ staff entitled, "READ Model Management Control Procedures," was sent to the committee members before the second meeting.

Second Meeting

The second meeting was held on December 8, 1978. Several review members raised the question of the relevance of forecasting in their written comments prepared for the second meeting. While this is an interesting philosophical question, the READ staff thought it was beyond the scope of the purpose of the review. The staff had the impression the review was concerned with a specific evaluation of the READ model developmental effort. The introduction of this issue raised the much broader question of whether EIA should respond to congrsssional requests or inquiries from other offices for analysis of regional policy impacts using models like READ. Since there was no opportunity to discuss this issue at the meeting, a paper by W. Rostow, "Energy, Full Employment, and Regional Development," containing a concise non-econometric statement of the importance of engaging in the type of regional analysis envisioned for the READ model was mailed to the review committee members after the meeting. Questions were also raised concerning the legal mandate for EIA to engage in regional analysis. While legislation does mandate that DOE engage in regional analysis, the role of EIA is uncertain. The committee members were asked to answer a number of specific questions for the third review meeting by Dr. C. Roger Glassey, Assistant Administrator, Applied Analysis, related to future regional modeling development in EIA. The general nature of the questions concerned the strength of the economy-energy interaction, the need for comprehensive models, and advice on alternatives to READ.

Question 1: Can the use of energy system variables as explanatory variables for projecting demographic and economic activity variables be dismissed in principle?

Question 2: What tests should be undertaken to investigate the strength of the energy system/regional demographic and economic activity relationships?

Question 3: How should EIA proceed to model and project energy system/regional demographic and economic activity relationships?

There were two problems that were to be discussed in the reviewers' response. How should EIA analyze the socioeconomic and environmental impacts of deregulation of crude oil prices and a moriatorium on nuclear power plants?

During the meeting, a number of reviewers proposed that a viable procedure would be to hire a regional energy economist for each region to do the analysis on a case by case basis. Several reviewers and the READ staff objected to this alternative, since the results could not be replicated and may be inconsistent with previous studies and data bases.

At the suggestion of Harvey Wagner, the READ staff was given an opportunity to respond in writing to the comments of the committee before the third meeting. The READ staff response contained a modification in the model development plan and corrections of a number of misconceptions of the reviewers concerning the characteristics of the data base and the model. The modifications included a proposal to estimate the model at an SMSA and remainder of State area rather than the county area.

Third Meeting

The third meeting was held on February 23, 1979. The first phase was a presentation of the SMSA proposal. The second phase was a general discussion of the responses and recommendations of the committee. The written recommendations were diverse and it was difficult to find a consensus on all points by the committee. The third and final phase, each reviewer was asked to summarize his position, including modifications of their earlier written statements at the conclusion of the meeting.

After the third meeting, the focus of attention shifted from the technical adequacy of the model to the cost and benefits of developing the model in relation to its projected uses and compared to alternative analytical procedures.

SMSA READ Model

The review process permitted us to refer back to the day to day modeling activity and allowed us to review the original plan of dividing the model development into three phases: Preliminary, full, and extension of the full model to include an environmental sector. The primary purpose for developing a preliminary model using OLS was to demonstrate its feasibility. Once feasibility was obtained, the full model would be constructed using information obtained from the prototype model and estimated with full information maximum likelihood techniques. It was anticipated that errors which were detected in constructing the smaller model would be avoided during construction of the larger model. In retrospect, this process contains internal inconsistencies as Jerry Hausman has indicated.

Feasibility can only be achieved if the appropriate estimation techniques are utilized so that the sign and magnitude of the regression coefficients will be free of major biases.

In addition, since the preliminary model utilized data, the resource management design was not as high in quality as that in the full scale model, the equations and forecasts would be subjected to justified criticism. Thus, it was decided to abandoned the original model development schedule and plan to estimate and simulate a model at the SMSA level. The remainder of this paper reviews the advantages of an SMSA version of the model in three areas: Observational unit, data availability, and modified resource management plan.

Observational unit.—The review committee was asked to comment on whether the county level geographical detail of READ is necessary to achieve its stated goals in their written report for the second meeting. The responses on the proper data unit were varied from using individual data, counties, SMSA, BEA, and States to the 10 Federal regions.

The major objections to using the county as the observational unit were quality of data and county as a self contained economic area. The READ review committee offered some very persuasive evidence that the READ staff had been optimistic about estimating the model in this fashion. The primary advantage of the county unit was the facility it offered in aggregating to other regional areas such as BEA, SMSA, etc. However, we have been convinced that the statistical problems associated with the data and estimation procedures outweigh the value of this aggregation flexibility.

The use of the SMSA region greatly reduces the situs adjustment problem that existed in the county data. Since the SMSA consists of contiguous counties of major economically interdependent areas they will incorporate the major retail and labor markets within a region. In addition, the regional differences between places of employment and residence will significantly decline using the SMSA region as an alternative to the county.

Data availability.—Census data provides a large amount of detailed information at 5-year intervals that is reported at the county level (10 years for housing). It should be noted that there is a disclosure problem with some detailed information at the county level. The annual surveys provide more aggregated data at SMSA, major SMSA, States, or large counties. The 2-digit SIC information is adequate for the industrial classification schemes of the preliminary READ model. The original plan was to use row adjustment scheme (RAS) techniques, utilizing the census county data and the annual SMSA survey data, to derive more accurate county level estimates for the full scale model. In addition, an effort was planned to utilize other data sources, primarily from regulatory agencies to supplement the data.

The new proposal was to directly utilize the SMSA information from the census surveys of agriculture, mining, manufacturing, housing, and government, as well as trade association data in the estimation of the model. This procedure has several advantages that should respond to the two major criticisms (derived data and appropriate region for analysis) made during the second review meeting. The amount of derived data will be significantly reduced particularly in output, equipment investment, personal consumption expenditures, and the receipts and expenditures of State and local government. There will still be some derived data, but this problem always occurs in regional analysis. The data for an SMSA region is also superior to the BEA region, since the Federal agencies, including census, do not collect data at the BEA regional level which must be aggregated from county data. It is also our contention that any viable analytic alternative will have to utilize a data base, software system, estimation routines, and simulation program similar to READ. This point was stressed in the third meeting by the READ staff and apparently began to receive acceptance by the committee by the end of the meeting.

Resource management plan.—The original resource management plan divided the model development effort into three phases: Preliminary, full, and addition of an environmental sector. The revised plan eliminated the division between the preliminary and full model. Harvey Wagner suggested that a viable development strategy would be to build a prototype model for 6 to 12 regions, before expanding to a large regional model. The READ staff suggested an alternative strategy of reducing the initial size of the model by reducing the number of equations, but keeping the number of regions at the SMSA and remainder of State areas. The equations would be estimated in successive stages as they are required to interface with the SEED models as they are completed. The first set of equations would be used to drive the residential sector, followed by the commercial, transportation, and industrial sectors. This alternative proposal has the advantage of the ease of software and forecast validation implicit in the Wagner proposal, while retaining the cross section time series data base required for efficient estimation.

Committee's Final Response

The final response of the committee at the end of the third meeting to the SMSA proposal, and the requirement for READ type analysis for driving energy demand models and for use in impact analysis, was not a unanimous recommendation to proceed or stop work on the modeling effort. Indeed, the comments ranged from disappointment at abandoning the county model, to the other extreme of David Freedman who advised that the model should be canceled. The READ staff stressed throughout the review process that the primary question was not designing and constructing a perfect model, but the allocation of scarce resources to achieve the goals as outlined in the planning phase.

David Brillinger and Leon Cooper stated that the county level model was being abandoned for the SMSA model at too early a stage. They both noted that all the criticism of the derived data were opinions, that should be tested empirically. Brillinger stressed the analogy of the use of derived data in READ, to his previous efforts of projecting election returns for States from key precincts. Brillinger also suggested that the equations should be estimated at the regional level where real data for the dependent variable exists. This would result in a model that is estimated using a hybrid set of regions including SMSA, State, and county. Cooper does not believe that highly aggregative models can answer the major energy policy questions satisfactorily.

In reference to READ, in his comments for the third meeting, he stated that,

> "... regarded it as a valuable evolutionary tool, to be used to also define data needs and to be modified with new data, studies, and results. As such, I continue to regard it as valuable and possibly of potentially greater value than conventional sources of such 'projections' and 'forecasts' .... I see no compelling reasons that a model that has fewer 'regions' and fewer variables and which must ultimately be tested in the same way as the READ model, will produce better results. I have every confidence that, in the long run, it will be far more inadequate than READ."

The comments of the other reviewers were not as concise as that of Brillinger, Cooper, and Freedman. Robert Dorfman has consistently maintained throughout the review process that,

> "The relations among major demographic, economic, and energy market variables are so intricate that a formalized system of equations is needed to keep track of them and produce a coherent view of what the future may bring forth. Without the discipline of such a set of equations, we are in danger of basing our planning upon impossible contingencies. It is probably unncessary to add the qualification that sets of equations have no common sense, especially in a world where structural relationships can change without notice, because of legislation or other influences outside the model. Therefore, model forecasts are only one ingredient of an informed judgment, albeit an import one."

His major concerns were the using of OLS for the preliminary model, the use of the county as the observational unit, and the theortical specification of several equations of the model. He stated that the SMSA proposal satisfies 90 percent of the objections that he had with the model.

Jerry Hausman was concerned with the consistency problem between the driving variables used in the current MEFS and the economic impact analysis undertaken using the solution of MEFS. He believed the major usefulness of READ would be to ensure accounting consistency between the inputs and outputs of MEFS including the incorporation of energy and economy feedbacks. He was concerned, however, that it may be at least 2 years before high quality energy data is available that could be used to support an integrated MEFS-READ system.

Harvey Wagner's major suggestion on the SMSA model was model development strategy, rather than model content. He also stressed that noneconometric methodologies may be more appropriate for forecasting the location of large energy construction projects, because of their discrete nature and absence of a time series of events. The READ staff was in agreement with this recommendation, and the CLDS system and other noneconometric analytical techniques were planned to be used for forecasting the completion of large discrete projects.

David Freedman indicated in his paper for the third meeting that he does not believe that data is of high enough quality to warrant further expenditure of resources on development of the model. Karen Polenski withdrew from the review after the second meeting because she had submitted a contract proposal to DOE that could conflict with her participation in the review.

While there was a diversity of opinion on a course of recommended action for EIA in regional modeling, a major conclusion preceived by the READ staff and concurred by C. Roger Glassey, Assistant Administrator of Applied Analysis, was the revised SMSA model proposal satisfactorily answered most of the technical questions raised in the first two review meetings. While the model is not perfect, the data base, software system, estimation procedure, and simulation routines are sufficiently flexible to be utilized in an SMSA READ model that could be used to drive the SEED models and for use in economic impact analysis.

CONCLUSION

This paper has attempted to document the managerial procedures used to plan, develop, and implement the READ model. Much of the material used in the paper was derived from original sources including FEA and DOE memos describing the past activities of the old DAD and present EUA offices. Thus, the paper can also serve as a partial historical record of the development of energy demand analysis in EIA. The two most prominent characteristics of the environment during this process, from the author's viewpoint, were repeated organizational changes and a conflict situation between devoting resources to short turnaround analysis or long-range upgrading of EIA analytical capabilities.

The constant change in the organizational hierarchy had the disruptive effect on the READ staff of recurring goal reorientation. Resources, which normally would have been devoted to completing models or upgrading existing models, were devoted to restructuring the organization and to rejustification of current activities. The formation of DOE and the current structure of EIA and AA should reduce this problem in the future.

The conflict situation over short-run versus long-run projects is more serious for the development of improved analytical capability in AA. The organizational structure of AA in conjunction with the previous short tenure of senior staff, has resulted in an environment where it is more rational to perform short-term analysis at the expense of long-range projects. While the importance of long-run analysis has always been recognized, the organizational structure for its implementation has never been successful. There are currently three procedures utilized in AA for developing long-range analytical capability:

- Contract research,
- Joint assignments of short-run and long-run duration projects within an organizational group, and,
- Creation of a separate research subgroup within an office.

Each research development process has its own virtues and deficiencies. Contract research has the supposed advantage that it permits DOE staff to be substituted by contract funds, thus, permitting an office to expand its functions. While this expansion is possible, it should be recognized that it is costly. Contracts must be monitored if they are to generate a useful product. The discussion of the difficulties that the READ staff encountered in processing contracts, has been experienced by other offices in AA. Careful contract monitoring does not provide sufficient knowledge of a model to enable the contract monitor to immediately utilize the model for analysis or to integrate it into a large modeling system when it is received from the contractor. A serious long-run result of using contracts to replace inhouse research is that the skills of the AA staff become more orientated toward contract administration rather than analysis.

Joint assignment of short-run and long-run analysis within an organizational group is an ineffective structure for long-run development efforts. This was the organizational structure used for the READ project from January 1976 to November 1977, when the staff was in DAD. This procedure places too much pressure upon organizational directors, who have always succumbed to short-term goal satisfaction by transferring resources away from short-term projects to long-term projects.

The creation of the READ in the EUA office in November 1977 eliminated the excesses of the earlier system. Unfortunately, the resources were not available within AA to complete the SMSA READ model, and other development efforts during 1979. Thus, the third organizational structure for successfully devoting a long-term commitment of resources to developing improved analytical procedures has not been empirically tested.

ACKNOWLEDGMENTS

The author wishes to thank B.D. Hong, Terry Morlan, and Michael Tannen for providing historical information that was used in the paper. Terry Morlan and Michael Tannen also provided valuable advice after reading earlier drafts. The careful typing and proofreading of Pat Francis and Helen Taylor is highly appreciated. Naturally any errors or omissions are the responsibility of the author. This paper represents the opinions of the author and does not represent the official views of the Energy Information Administration, Department of Energy.

Appendix A. **Review Committee Consultations**

David Brillinger
University of California, Berkeley

Leon Cooper
Southern Methodist University

Robert Dorfman
Harvard University

David Freedman
University of California, Berkeley

Jerry Hausman
Massachusetts Institute of Technology

J. Daniel Kazzom
Institute for Research in Energy and Economic Modeling
San Francisco, California

Karen Polenski
Massachusetts Institute of Technology

Harvey Wagner
University of North Carolina

Basic-Nonbasic Allocation Error and Least Squares Bias in Regional Export Base Models

Boyd L. Fjeldsted
University of Utah

Perhaps the most widely used approach in the construction of small area forecasting models is the export base scheme. This approach assumes that total economic activity in an appropriately defined geographical area or region may be partitioned into two components:

1. An export or basic activity component.
2. A local service or residentiary activity component.

The approach further assumes that the export activity component is exogenously determined, while the residentiary activity component is endogenously determined.[1]

The simplest and most primitive form of the export base approach involves the calculation of a basic-service ratio from a single data point with the presumption that the ratio remains stable, at least with respect to any contemplated changes in the level of basic activity. Other methods have now largely superseded the simple basic-service ratio technique. The assumption of strict proportionality between basic activity and residentiary activity is unnecessarily restrictive, while the use of a single data point results in unacceptable large sampling variability.

The method of ordinary least squares regression analysis (OLS) is one of the most commonly employed alternatives to the simple basic-service ratio technique. And though Park (see reference 5), has shown that the OLS estimators of the parameters are critically sensitive to the allocation coefficient vector used to distribute industrial activity between the basic and nonbasic categories, little has been done in the way of formally deducing the nature of the bias under various allocation error circumstances.

If all sectors within the industrial classification scheme employed by data providers were purely basic or purely residentiary the export base model builders would not have to be concerned with allocation error, except for the special case of basic-nonbasic classification error. No matter how detailed the industrial classification system, the fact remains that most sectors must be regarded as "mixed." Even at the individual establishment level, it is commonly the case that some activity is for export purposes while other activity is for local service purposes.

The purpose of this paper is to consider the statistical implications of various types of allocation error, both systematic and random. Among the principal conclusions are the following propositions:

1. Purely random allocation error results in the OLS estimator of the multiplier parameter being biased toward minus one (instead of toward zero as in the classical error-in-variables situation).

2. Systematic allocation of residentiary activity to the basic category results in a downward bias in the OLS estimator of the multiplier parameter.

3. Systematic allocation of basic activity to the residentiary category ordinarily can be expected to lead to an upward bias in the OLS estimator of the multiplier parameter.

4. Allocation of activity according to the so-called "location quotient" approach is likely to result in an upward bias in the OLS estimator of the multiplier parameter, though, paradoxically, if the variance in the regressor is small a downward bias could result.

Finally, it is observed that the allocation error problem described here is really not peculiar to regional export base models. Generally, the presence of an endogenous error component in a nominally specified exogenous variable (or vice versa) is a pervasive source of specification error in econometric model construction.

Stochastic Allocation Error

Suppose that residentiary activity Y is a linear function of basic activity X, subject to a random disturbance σ:

$$Y = \alpha + \beta X + \delta \ .$$

Further suppose that the true values of these variables are never observed, due to stochastic error in the allocation of total economic activity between basic and residentiary activity. When Y is overestimated, X is underestimated by an equal amount, and vice versa.

Let Y* denote the estimated value of Y, and X* the estimated value of X.

Then:

$$Y^* = Y + \varepsilon$$

and

$$X^* = X - \varepsilon \quad ,$$

where ϵ is a stochastic error variable. The OLS estimator β would be given by the expression:

$$\hat{\beta} = \frac{\operatorname{cov}(X^*, Y^*)}{\operatorname{var}(X^*)}$$

where var and cov are used to denote the sample variance and covariance, respectively.

By substitution:

$$\hat{\beta} = \frac{\operatorname{cov}(X-\varepsilon,\ Y+\varepsilon)}{\operatorname{var}(X-\varepsilon)} \ .$$

But since the covariance of two sums is equal to the sum of the covariances of the individual terms, and the variance of a sum

[1] Most export base models have used employment as the measure of economic activity primarily because employment data historically have been more readily available for small areas than data on other suggested measures of activity, such as income or earnings.

is equal to the sum of the variances plus twice the covariance it follows that

$$\hat{\beta} = \frac{\begin{array}{r}\mathrm{cov}(X,Y) + \mathrm{cov}(X,\varepsilon) \\ + \mathrm{cov}(Y,-\varepsilon) + \mathrm{cov}(\varepsilon,-\varepsilon)\end{array}}{\mathrm{var}(X) + \mathrm{var}(-\varepsilon) + 2\mathrm{cov}(X,-\varepsilon)}$$

By substituting for Y, and rearranging signs, it can be seen that

$$\hat{\beta} = \frac{\begin{array}{r}\mathrm{cov}(X,\beta X+\delta) + \mathrm{cov}(X,\varepsilon) \\ - \mathrm{cov}(\beta X+\delta,\varepsilon) - \mathrm{cov}(\varepsilon,\varepsilon)\end{array}}{\mathrm{var}(X) + \mathrm{var}(\varepsilon) - 2\mathrm{cov}(X,\varepsilon)}$$

$$= \frac{\begin{array}{r}\mathrm{cov}(X,\beta X) + \mathrm{cov}(X,\delta) + \mathrm{cov}(X,\varepsilon) \\ - \mathrm{cov}(\beta X,\varepsilon) - \mathrm{cov}(\delta,\varepsilon) - \mathrm{cov}(\varepsilon,\varepsilon)\end{array}}{\mathrm{var}(X) + \mathrm{var}(\varepsilon) - 2\mathrm{cov}(X,\varepsilon)}$$

Assuming X, σ and ϵ to be independent, it follows from the Slutsky Theorem on probability limits that:

$$\mathrm{plim}\ \hat{\beta} = \frac{\beta\sigma_{XX} - \sigma_{\varepsilon\varepsilon}}{\sigma_{XX} + \sigma_{\varepsilon\varepsilon}} \quad ,$$

where $^{\sigma}XX$ denotes the probability limit of var (X), and $^{\sigma}\epsilon\epsilon$ denotes the probability limit of var (ϵ).[2] By further rearrangement,

$$\mathrm{plim}\ \hat{\beta} = \frac{\beta - \dfrac{\sigma_{\varepsilon\varepsilon}}{\sigma_{XX}}}{1 + \dfrac{\sigma_{\varepsilon\varepsilon}}{\sigma_{XX}}} \quad .$$

Thus it can be seen that the asymptotic bias in $\hat{\beta}$ depends critically on the variance ratio $\sigma_{\epsilon\epsilon}/\sigma_{XX}$. Denoting this variance ratio by θ, the probability limit of $\hat{\beta}$ may be expressed as

$$\mathrm{plim}\ \hat{\beta} = \frac{\beta - \theta}{1 + \theta} \quad .$$

Note that since $\theta > 0$, for $\beta > 0$ plim $\hat{\beta} < \beta$, i.e., the asymptotic expectation of $\hat{\beta}$ is smaller than β. Note also that:

$$\lim_{\theta\to\infty} \mathrm{plim}\ \hat{\beta} = -1 \quad ,$$

i.e., the effect of random allocation error is to bias the least squares estimator toward minus one. This contrasts with the ordinary errors-in-variables situation, where the effect is to bias the least squares estimator toward zero.[3] But it may also be observed that

$$\lim_{\theta\to 0} \hat{\beta} = \beta \quad .$$

This means that, as in the ordinary errors-in-variables case, if the variance in X is large enough relative to the error variance, the asymptotic bias in $\hat{\beta}$ becomes negligible.

Symstematic Allocation of Residentiary Activity to the Basic Sector

Suppose that residentiary activity Y has two components, a principal component Y*, and a secondary component Y** that is erroneously allocated to the basic activity sector. Activity in each residentiary sector is a linear function of basic activity X, subject to a random disturbance, i.e.,

$$Y^{*} = \alpha_1 + \beta_1 X + \zeta$$

and

$$Y^{**} = \alpha_2 + \beta_2 X + \xi \ ,$$

where ζ and ξ are random disturbances. Summing the two equations gives

$$Y^{*} + Y^{**} = (\alpha_1+\alpha_2) + (\beta_1+\beta_2)X + (\zeta+\xi)$$

which may be expressed simply as

$$Y = \alpha + \beta X + (\zeta+\xi) \quad ,$$

where $\alpha = \alpha_1 + \alpha_2$ and $\beta = \beta_1 + \beta_2$. The true export base multiplier parameter is of course β, which reflects the actual effect of basic activity X on nonbasic activity Y.

However in estimating β, the secondary residentiary component Y** is systematically (but erroneously) allocated to the basic activity sector. The OLS estimator of $\hat{\beta}$ is thus obtained by regressing Y* on X + Y**, i.e.

$$\hat{\beta} = \frac{\mathrm{cov}(X+Y^{**},Y^{*})}{\mathrm{var}(X+Y^{**})}$$

$$= \frac{\mathrm{cov}(X,Y^{*}) + \mathrm{cov}(Y^{**},Y^{*})}{\mathrm{var}(X) + \mathrm{var}(Y^{**}) + 2\mathrm{cov}(X,Y^{**})}$$

$$= \frac{\mathrm{cov}(X,\beta_1 X+\zeta) + \mathrm{cov}(\beta_2 X+\xi,\beta_1+\zeta)}{\mathrm{var}(X) + \mathrm{var}(\beta_2 X+\xi) + 2\mathrm{cov}(X,\beta_2 X+\xi)}$$

$$= \frac{\begin{array}{r}\mathrm{cov}(X,\beta_1 X) + \mathrm{cov}(X,\zeta) + \mathrm{cov}(\beta_2,\beta_1 X) \\ + \mathrm{cov}(\beta_2,\zeta) + \mathrm{cov}(\beta_1,\xi) + \mathrm{cov}(\zeta,\xi)\end{array}}{\begin{array}{l}\mathrm{var}(X) + \mathrm{var}(\beta_2 X) + \mathrm{var}(\xi) \\ + 2\mathrm{cov}(\beta_2 X,\xi) + 2\mathrm{cov}(X,\beta_2 X) + 2\mathrm{cov}(X,\xi)\end{array}}$$

Assuming X to be independent of ζ and ξ ,

$$\mathrm{plim}\ \hat{\beta} = \frac{\beta_1\sigma_{XX} + \beta_1\beta_2\sigma_{XX} + \sigma_{\zeta\xi}}{\sigma_{XX} + \beta_2^2\sigma_{XX} + \sigma_{\xi\xi} + 2\beta_2\sigma_{XX}} \quad ,$$

[2] Goldberger, Arthur S., *Econometric Theory*, New York: John Wiley and Sons, 1964, pp. 118-119 provides a statement and discussion of the Slutsky Theorem on probability limits.

[3] See, e.g., Theil, Henri, *Principales of Econometrics*, New York: John Wiley and Sons, 1971, pp. 608-609.

where σ_{XX} denotes the probability limit of var(X), $\sigma_{\zeta\xi}$ denotes the probability limit of cov (ζ,ξ), and $\sigma_{\xi\xi}$ denotes the probability limit of var(ξ). Rearranging terms gives

$$\text{plim } \hat{\beta} = \frac{\beta_1(1+\beta_2) + \dfrac{\sigma_{\zeta\xi}}{\sigma_{XX}}}{(1+\beta_2)^2 + \dfrac{\sigma_{\xi\xi}}{\sigma_{XX}}} \quad .$$

It is clear that $\hat{\beta}$ is not an unbiased estimator of the export base multiplier parameter β, and in fact the asymptotic expectation of $\hat{\beta}$ may be either larger or smaller than β, depending on the relative magnitudes of β_1, β_2, σ_{XX}, $\sigma_{\zeta\xi}$ and $\sigma_{\xi\xi}$.

The probability limit of $\hat{\beta}$ may alternatively be expressed as

$$\text{plim } \hat{\beta} = \frac{\beta_1(1+\beta_2) + \rho}{(1+\beta_2)^2 + \phi} \quad ,$$

where ρ denotes the ratio of $\sigma_{\zeta\xi}$ to σ_{XX}, and ϕ denotes the ratio of $\sigma_{\xi\xi}$ to σ_{XX}. Note that as $\sigma_{\xi\xi}$ approaches zero, $\sigma_{\zeta\xi}$ approaches zero. Therefore, as ϕ approaches zero, ρ also approaches zero. Thus,

$$\lim_{\phi \to 0} \text{plim } \hat{\beta} = \frac{\beta_1(1+\beta_2)}{(1+\beta_2)^2} = \frac{\beta_1}{1+\beta_2} \quad .$$

Note that if the multiplier coefficients β_1 and β_2 are greater than zero,

$$\frac{\beta_1}{1+\beta_2} < \beta_1 + \beta_2, \text{ i.e.,}$$

When the disturbance variances are small relative to the variance of the regressor, plim $\hat{\beta} < \beta$. This means that when the regressor contains a nonbasic component, the OLS estimator of the export base multiplier possesses a downward bias (as long as the variances of disturbances are small relative to the variance in the regressor).

Systematic Allocation of Basic Activity to the Residentiary Sector

Suppose that basic activity X consists of two components, a principal component X* and a secondary component X** and that residentiary activity Y is a linear function of activity in the two basic sectors, subject to a random disturbance δ, i.e.,

$$Y = \alpha + \beta X^* + \gamma X^{**} + \delta \quad .$$

Suppose that in estimating the principal basic sector multiplier parameter β, the secondary basic component X** is systematically allocated to the residentiary sector. The OLS estimator of β would then be given by the expression:

$$\hat{\beta} = \frac{\text{cov}(X^*,Y+X^{**})}{\text{var }(X^*)}$$

$$= \frac{\text{cov}(X^*,Y) + \text{cov}(X^*,X^{**})}{\text{var}(X^*)}$$

$$= \frac{\text{cov}(X^*,\beta X^*+\gamma X^{**}+\delta) + \text{cov}(X^*,X^{**})}{\text{var}(X^*)}$$

$$= \frac{\text{cov}(X^*,\beta X^*) + \text{cov}(X^*,\gamma X^{**}) + \text{cov}(X^*,\delta) + \text{cov}(X^*,X^{**})}{\text{var}(X^*)}$$

$$= \frac{\beta\text{var}(X^*) + (1+\gamma)\text{cov}(X^*,X^{**}) + \text{cov}(X^*,\delta)}{\text{var}(X^*)}$$

Assuming X* and δ to be independent,

$$\text{plim } \hat{\beta} = \beta + (1+\gamma)\frac{\sigma_{X^*X^{**}}}{\sigma_{X^*X^*}} \quad ,$$

where $\sigma_{X^*X^{**}}$ denotes the probability limit of cov(X*,X**) and $\sigma_{X^*X^*}$ denotes the probability limit of var(X*). In general then, the erroneous allocation of a basic sector component to the nonbasic sector will result in a biased estimator of the export base multiplier for the basic sector remaining as the regressor. In the usual case, where common resource base, agglomeration or interindustry requirements phenomena are present, the covariance of activity in the primary basic sector and activity in the secondary basic sector will be positive. Thus, the OLS estimator of the export base multiplier parameter β may ordinarily be expected to have an upward bias when an erroneous allocation of a basic sector component to the nonbasic sector has occurred.[4]

Allocation of Activity by Means of the Location Quotient Technique

One of the most popular devices for allocating total activity between its basic and residentiary components is the so-called "location quotient" technique. This method assumes that the regional residentiary requirement for any sector is proportional to total regional activity, with the constant of proportionality being equal to the ratio of national activity in the sector to total national activity. The location quotient method implicitly

[4] Polzin, Paul E., "Urban Employment Models: Estimation and Interpretation," *Land Economics*, 49, (1973) p. 232; has argued that since basic sector activity is exogenous, the covariance of activity in one basic sector and activity in any other basic sector should be equal to zero. On the basis of this argument he concludes that to avoid bias in estimating the export base multiplier it is permissible to include some basic activity in the residentiary sector as long as the designated basic activity regressor does not include an endogenous activity component. However, Polzin's argument is fallacious. It is a basic tenet of statistical inference that correlation between variables does not imply that they are causally related. Or to express the idea slightly differently—the absence of a causal relationship between variables does not imply an absense of correlation between them. In fact, correlation among exogenous variables is just the commonplace multicollinearity phenomenon, quite generally characteristic of large-scale econometric models.

assumes that the national economy is closed and stationary. All activity within the national economy is directed toward the satisfaction of residentiary requirements. It is implicitly assumed that any regional mix of residentiary requirements is the same as the national mix, which implies no regional differences in consumer preferences or technology. Under these assumptions the regional mix of residentiary *requirements* by industrial sector will be the same as the national *activity* mix by sector.

The location quotient for a sector is defined as the proportion of total regional activity accounted for by the sector divided by the proportion of total national activity accounted for by the sector. Under the previous assumptions the location quotient provides an estimate of the ratio of total activity in a sector to residentiary requirements for the sector. If the location quotient exceeds unity it is presumed that the excess of activity over residentiary requirements is exported. If the location quotient falls short of unity it is presumed that the deficit in activity is satisfied through imports.

These notions may be expressed mathematically as follows: Let Y^*_{ij} denote the estimated residentiary requirement for sector i in region j and T_{ij} denote total activity in sector i in region j. Then according to the location quotient approach,

$$Y^*_{ij} = \frac{\sum_j T_{ij}}{\sum_i \sum_j T_{ij}} \quad \sum_i T_{ij}$$

where the range of summation for index j extends over all regions in the nation. Estimated export activity for sector i in region j is then given by the expression

$$X^*_{ij} = \begin{cases} T_{ij} - Y^*_{ij} & \text{for } T_{ij} \geq Y^*_{ij} \\ 0 & \text{otherwise.} \end{cases}$$

That the location quotient technique results in a systematic overestimation of residentiary activity and consequent underestimation of basic activity has been pointed out by Blumenfeld (see reference 1, pp. 119-120), Leven (see reference 4, p. 255), Tiebout (see reference 9, pp. 48-49), Pratt (see reference 7, p. 122), and Greytak (see reference 2, p. 388), *inter alios*. When total activity in a sector exceeds the estimated residentiary requirement, the excess is assumed exported. However, if the residentiary requirement for the sector is partially satisfied through imports, then actual residentiary activity will be smaller than the residentiary requirement and actual export activity will be larger than estimated export activity by the equivalent of the amount of imports. Similarly, when total activity in a sector falls short of the estimated residentiary requirement, the deficit is assumed to be satisfied through imports, and export activity is estimated to be zero. But if actual imports exceed the deficit, then actual export activity will be greater than zero by the equivalent of the "excess" imports. In either case residentiary activity is overestimated and export activity is underestimated by the equivalent of the implicit underestimation of imports.

In fact, the occurrence of exports in an industrial sector does not preclude the occurrence of imports, and vice versa. Even the level of detail industrial sectors are defined so broadly as to include distinctly different product classes. But with complete product homogeneity within industrial sectors, exports and imports of the same product would occur due to the fact that regions encompass a set of geographical points rather than a single point. Thus, exports may occur at one boundary point, while imports are occurring at another boundary point. It should be noted that cross-hauling of physically indistinguishable products occur because of various institutional arrangements and market "imperfections."

When the location quotient technique is used in conjunction with a single data point, the systematic overstatement of residentiary activity and understatment of basic activity obviously results in an upward bias in the estimated export base multiplier. When the technique is used in conjunction with the method of ordinary least squares regression analysis, the nature of the resulting bias is less obvious. Restricting the discussion to a single region and dropping the regional subscript, let ϵ_i denote the overestimate of residentiary activity in sector i obtained by using the location quotient technique, i.e.,

$$\varepsilon_i = Y^*_i - Y_i \ ,$$

where Y^*_i denotes the residentiary requirement in the i^{th} sector estimated according to the location quotient technique, and Y_i denotes actual residentiary activity in the i^{th} industrial sector.

It follows the ϵ_i equals the underestimate of export activity in the i^{th} sector, i.e.,

$$\varepsilon_i = X_i - X^*_i \ ,$$

where X_i denotes actual export activity in the i^{th} sector and X^*_i denotes export activity in the i^{th} sector estimated according to the location quotient technique.

Note that sector i implicit estimated imports M^*_i would be given by the expression:

$$M^*_i = \begin{cases} Y^*_i - T_i & \text{for } T_i < Y^*_i \\ 0 & \text{otherwise.} \end{cases}$$

Since exports in every sector are explicitly underestimated by the amount that imports are implicitly underestimated, it follows that ϵ_i is also equal to the difference between actual imports and implicit estimated imports, i.e.,

$$\varepsilon_i = M_i - M^*_i \ ,$$

where M_i denotes actual imports in the i^{th} sector.

Rearranging the original error equations gives:

$$Y^*_i = Y_i + \varepsilon_i$$

and

$$X^*_i = X_i - \varepsilon_i \ .$$

Then summing over all sectors gives:

$$\sum_i Y_i^* = \sum_i Y_i + \sum_i \varepsilon_i$$

and

$$\sum_i X_i^* = \sum_i X_i - \sum_i \varepsilon_i \ .$$

Now letting Y* denote $\sum_i Y_i^*$, Y denote $\sum_i Y_i$, X* denote $\sum_i X_i^*$, X denote $\sum_i X_i$ and ϵ denote $\sum_i \epsilon_i$, the previous two equations may be written as:

$$Y^* = Y + \varepsilon$$

and

$$X^* = X - \varepsilon \ .$$

Thus the problem may be viewed as essentially similar to the random allocation problem discussed in Stochastic Allocation Error Section. That the error term ϵ may not assume negative values in the location quotient special case is a trivial distinction. However, the possibility of the error term being correlated with export activity precludes the vanishing of the covariance term involving X and ϵ.

Again, suppose that residentiary activity Y is a linear function of basic activity subject to a random disturbance δ:

$$Y = \alpha + \beta X + \delta \quad .$$

Now assuming independence between X and δ and between ϵ and δ (but not between X and ϵ), it follows that

$$\text{plim } \hat{\beta} = \frac{\beta\sigma_{XX} + (1-\beta)\sigma_{X\varepsilon} - \sigma_{\varepsilon\varepsilon}}{\sigma_{XX} - 2\sigma_{X\varepsilon} + \sigma_{\varepsilon\varepsilon}} \quad ,$$

where σ_{XX} denotes the probability limit of the variance of X, $\sigma_{X\epsilon}$ denotes the probability limit of the covariance of X and ϵ, $\sigma_{\epsilon\epsilon}$ denotes the probability limit of the variance of ϵ.

Note that as σ_{XX} approaches zero, so does $\sigma_{X\epsilon}$, so that

$$\lim_{\sigma_{XX}\to 0} \text{plim } \hat{\beta} = \frac{-\sigma_{\varepsilon\varepsilon}}{\sigma_{\varepsilon\varepsilon}} = -1 \ .$$

Thus, if the variance of basic activity is "small" relative to the variance of the error term, the effect of allocation error resulting from the location quotient technique may be bias to the OLS estimator of the multiplier parameter downward. This could very well be the case if the sample consists of a limited number of time series data points.

On the other hand if the variance of basic activity is large enough relative to the error variance it could very well be the case that the probability limit of $\hat{\beta}$ exceeds β. In fact it can be argued that ordinarily the underestimate in export activity resulting from the location quotient technique might be expected to be roughly proportional to export activity itself, in which case $\sigma_{X\epsilon}$ would be positive. Especially in the situation where β is no greater than unity, the effect of a positive $\sigma_{X\epsilon}$ could more than offset the effect of $\sigma_{\epsilon\epsilon}$, resulting in an upward bias in the OLS estimator of the export base multiplier.

Concluding Observations

The type of problem treated here is by no means unique to regional export base models. The problem exists in any sort of econometric modeling context where there is the possibility that a nominally specified exogenous variable contains an endogenous error component or vice versa. This is no doubt the case in most regional econometric modeling efforts, where model specification is largely dependent upon data availability. But even large scale national econometric models should not be considered immune from the difficulty. Only in an ideal data world populated with ideal econometricians could the problem completely be eliminated.

In view of the apparent pervasiveness of the problem it is somewhat surprising that so little attention has been given to the estimating-error and forecasting-error implications of the problem. The arguments developed within the context of the application of least squares regression analysis to the estimation of the parameters of a regional export base model suggest that the bias resulting from exogenous-endogenous "impurity" in the variables may be considerable, and that the magnitude of consequential forecasting error may be larger than hitherto realized.

REFERENCES

1. Blumenfeld, Hans, 'The Economic Base of the Metropolis," *Journal of the American Institute of Planners*, 21 (1955).

2. Greytak, David, "A Statistical Analysis of Regional Export Estimating Techniques," *Journal of Regional Science*, 9 (1969).

3. Goldberger, Arthur S., *Econometric Theory*, New York: John Wiley and Sons, 1964.

4. Leven, Charles L., "Measuring the Economic Base," *Papers and Proceedings of the Regional Science Association*, 2 (1956).

5. Park, Se-Hark, "Least Squares Estimates of the Regional Employment Multiplier: An Appraisal," *Journal of Regional Science*, 10 (1970).

6. Polzin, Paul E., "Urban Employment Models: Estimation and Interpretation," *Land Economics*, 49 (1973).

7. Pratt, Richard T., "An Appraisal of the Minimum Requirements Technique," *Economic Geography*, 44 (1968).

8. Theil, Henri, *Principles of Econometrics*, New York: John Wiley and Sons, 1971.

9. Tiebout, Charles M., *The Community Economic Base Study*, Supplementary Paper No. 16, Committee for Economic Development, December 1962.

Discussant

Joseph W. Duncan
Department of Commerce

Can we. forecast small area data? The three different papers today have taken totally different approaches to answering that question.

The first paper, "A Model of Construction Activity in Subnational Areas" gave an empirical view of the question. The second presentation, "Developing and Managing a Small Area Forecasting Model—READ" gave a management view of the question. The third paper, "Basic-Nonbasic Allocation Error and Least Squares Bias in Regional Export Base Models" gave a mathematical view of the question.

Given the diversity of those three papers and approaches, it might be difficult to develop a common theme. However, I believe there is a common theme that is of significant long term interest to all of us who have an interest in small area forecasting. That theme relates to the problems in specifying the data base for small area forecasting. Frank Hopkins has a more optimistic view than I do about the availability of data that will help us meet our needs. I will come back to that in a moment, but first I want to make some specific comments about the individual papers.

The first paper on construction modeling presents a very interesting approach to looking at local labor markets. The author approaches the problem quite realistically in terms of the models that are used. However, there is not much information about the. specific data base and when one begins to think through the problems of data that are necessary for estimating construction sectors, other than residential construction—the example which is used, it appears that it will be difficult to actually estimate a number of sectors which are to be examined.

The logic of the models is clear and I think the presentation is quite constructive in presenting alternative estimating procedures. I did find it a little odd that the narrative statement leaves out migration. However, when you look at the actual equations' migration is there. Finally, I want to repeat what the author says because it relates to my general comments. The last paragraph states "applying good statistical methods to small area data is not easy. The data are often crude, and it is easy to be overwhelmed by the data management task. We are encouraged, that this exercise showed that sophisticated methods do result in different findings." That does leave the forecaster up in the air a little bit about what to do. But I think that is the essence of the problem.

I have been reasonably familiar with PIES model. It was a little odd to me that Hopkins took a position that the data are really not a problem.

There are many rich sources of data. Most of my daily working time is devoted to trying to improve these existing data bases. It is not my general impression that there is a large supply of "good" data. Quite frequently, the data that are available are not defined with adequate procedures of quality control on the concepts, definitions that are being used to collect the data, the quality of the collection technique itself, or with the specification and documentation of the adjusted data being made available. One can obtain many numbers, but there are many problems. For example, when we look at Federal/State systems of data where the data collection is under the control of very diverse units in the various States the available estimates are of differing quality. Much of the data that exist are weak, relative to accurate modeling procedures. When you add that to Fjeldsteds' comments about structural problems of specifying a regional model, it is easy to be pessimistic.

I can convert Fjeldsted's mathematical statements into data statements also. What he is saying is that first of all you have a model specification problem. There are problems of classifying activity into the basic two categories—export or residentiary. This is a data problem—which data are associated with which sector? Once you have set up the fundamental structure, there are problems of definition within the individual industrial sectors.

It's relatively easy to suggest that what we really need is an integrated set of regional economic data. We need a set of State economic accounts that will permit estimates for example, of flows, that is, imports/exports in the various States. The difficulty is, that we have a split personality at the Federal level with respect to local area data collection.

Starting with the new Federalism that former President Nixon introduced, there has been a view that the Federal responsibility is to collect and define the basic programs and then to allocate them back into the States for local administration where people have more understanding of specific local needs and where the differentials among localities can be taken into account. One needs to have comparable Federal/State data so that the allocations can be made equally and the policies can be evaluated. Instead, we have given to each State the responsibility to do as it will in the area of data collection.

The hard reality is, that there is no lobby for good local area information. Each local area tends to be very introspective in terms of its own limitations and resources. The differences among the areas are quite clear.

After reviewing these papers, I remain a little pessimistic about the Chairman's overall question "can we forecast small area data?" I do not think we can. Despite the sophistication of the estimating techniques or the models, our data bases are not up to the task. We are going to have to make some drastic improvements in the data base if we're going to be successful in meeting the challenge that was proposed to us—developing good models and good results empirically for forecasting small areas.

1980 Census— Small-Area Statistics Program

Introduction

Irving Roshwalb
Audits & Surveys, Inc.

The increasing role of the Census Bureau data in effecting social and economic programs has brought new pressure on the Census Bureau to produce more detailed, more varied, and more accurate reports. The papers presented here deal with the Bureau's plans to provide small area data, that is, data reported on a State level or below. That is not to slight the issue of accuracy which is of paramount importance but to leave that for other discussions.

In a recent article,[1] Keyfitz has estimated that "... one missing person in apportionment could deprive a State of a representative and of more than $100 in allocation of revenue sharing and other federal grants in a single year." The use of small area data to solve problems related to the distribution of federal funds, political representation, and to economic and social planning are among the better known applications of these data. Perhaps less known but certainly important are the use of the data in various aspects of business operations. Small area data are used extensively in such applications as retail store location, mapping sales routes, product sampling, sales analysis, test marketing, and, of course, sample selection for survey research.[2]

The community of users of the Census Bureau's small-area statistics is one of varied interests and applications. The papers by Turner and Garland provide a foretaste of the materials the Census Bureau plans to provide its user community. Marshall L. Turner, Assistant Chief of the Decennial Census Division of the Census Bureau, describes the small-area statistics program for the 1980 Census. In addition, he discusses some of the considerations that went into developing this program. The second paper is by Michael G. Garland, Chief, Data User Services Division of the Census Bureau. He has the task, which, if successful, will make the census effort, including the small-area statistics program, as worthwhile as possible. In his paper, he discusses the procedures that have been proposed to insure that the published data find their way off the Census shelf and *used*, rather than onto the shelf and merely *available*.

The discussants themselves come to the session from somewhat different disciplines and points of view. Dr. Pearl Kamer is Chief Economist of the Nassau-Suffolk Regional Planning Board as well as Adjunct Professor of Urban Public Policy at the State University at Stony Brook. Mr. Edward Spar is the President of Market Statistics, a company that makes extensive use of census data, and has had close associations with marketing applications of small-area statistics.

In addition to providing data for the apportionment of the House of Representatives and State legislature, small-area statistics guides expenditures on such programs as aids to mass transportation and water resources planning. The New York Times estimates that "no fewer than 107 Federal aid programs are based wholly or partly on population."[3] This underlines the wide use of small-area statistics and makes the papers presented here timely and valuable.

[1] Keyfitz, Nathan (1979, "Information and Allocation: Two Uses of The 1980 Census," *The American Statistician*, 33, pp. 45-50.

[2] Dutka, Solomon, Frankel, Lester R., and Roshwalb, Irving (1971), "A Marketer's Guide to Effective Use of 1970 Census Data." *Modern Marketing* Series, No. 8, Audits & Surveys, Inc., New York.

[3] New York Times, September 16, 1979. This article was published and located during the editing of these remarks.

1980 Small-Area Statistics Program

Marshall Turner
Bureau of the Census

I have been at the Census Bureau for about 15 years, and I have given a lot of presentations for various professional meetings and never missed being present at the session where I was to speak. I had a bit of fright last week when a call came from a lady in Chicago—University of Chicago—working on her doctrinal dissertation. She indicated she was most anxious to get certain statistics from the 1978 dress rehearsal census that we conducted in Richmond, Virginia. Only last week did these data go to the printer and they are not available yet. I indicated to her that she should go ahead and send us a note indicating why she needed the data. Perhaps we could make available to her a few prepublication figures for academic purposes.

During our conversation, I asked her to tell me why she needed the data. She said that in a couple of weeks she would be presenting a paper at the ASA. I said, "What a coincidence, so am I. I'll tell you what I can do, if necessary, I could even bring the data with me and I can meet you somewhere and discuss the data with you." She said fine, "I'll be staying across the river near Harvard." I said, "Harvard, for the ASA meetings?"

She persisted in trying to convince me that she really was going to give a paper at the ASA. And I said, "You know, this is very disturbing. For a long time I have been giving papers and I have never appeared in the wrong city, but it seems that either I have made a mistake or someone else has." Finally it came to both of us almost simultaneously that we had better define what we meant by ASA. Of course she was talking about the "sociological" meetings and I was talking about the "statistical" meetings. I almost went to Boston instead of here in Washington, D.C.

The planning process for the 1980 census has been going on formally since 1974 and informally since we produced our last publication from the 1970 census in 1973. The overall direction of that planning effort for 1980 has been influenced in the past years, and very markedly so, by the proliferation of social and economic programs at the Federal, State, and local level that effect all elements of our country, all jurisdictions, and by certain congressional actions. In getting ready for the 1980 census, the Census Bureau anticipated, or rather Dave Kaplan, (who has since retired) foresaw a lot of this and felt very stongly in the Census Bureau carrying through and seeking the widest possible source of input in getting ready for the 1980 census. But with the clear and stated caveat that we could not do all the things that everyone would ask us to do. As one of my colleagues at the Census Bureau says, if we had included all the hundreds of questions that had been recommended in the past 9 years, we would have a questionnaire that weighs about 25 pounds and would not fit into any of the mailboxes. We are planning to deliver the questionnaires on March 28, 1980. What we have tried to do is operate within the system of constraints, both in terms of respondent perception and understandability of the questions, and what we consider is a realistic appraisal of what the respondent is willing to bear in producing the types of data that we feel will be most needed in the decade of the 1980's. Needless to say, it will be a compromise—it won't meet everyone's needs.

In terms of the planning input, there are three major areas of small area data needs that we hear repeatedly:

1. The need for more precise total population counts for performing the function of reapportionment and redistricting. This means not only the drawing of congressional district boundaries but also the drawing of legislative district boundaries for the State bodies and at the sub-State level for city and county forms of representative government. This arose during the 1960's from decisions of the courts on one-person/one-vote.

2. The burgeoning number of Federal programs. There is about $50 billion annually in Federal funds that are allocated back to the 39,000 general purpose local governments in this country, and the allocations are based partly on census information. This permeates all jurisdictions in the country and many of you here represent practitioners of these planning programs. The sources of the recommendations are from local elected officials; especially with the advent of Federal revenue sharing in 1972, not to mention the funds provided under Title 1 of the Elementary and Secondary Education Act, CETA programs, the Community Development Block Grant programs, and so forth. They also come from members of congress. On the appropriation committee in the Senate, that oversees our budgetary request, we have Senator Patrick Leahy who represents many rural areas, many small towns in New England, and who feels very strongly about the data needs of these communities.

3. The private sector has strongly pursued the need for more census information, more small-area statistics for the purposes of the business sector of the country.

Let's talk briefly about some of the operational changes which the Census Bureau has undertaken to try to effect some of the small area data products that seem to be needed. For those of you familiar with the 1970 Census, the first set of tabulations represented about 400 cells of information summarized from the questions that were asked of every household in the nation in 1970. These data, and I'll only emphasize the smallest level for which they were tabulated, were produced for enumeration districts in the more rural, less densely settled areas of the country, and for block groups in the more urbanized or more densely settled portions. The third count computer summary tabulations that came about a year to a year and a half later, represented summations of the data collected on a 100 percent basis for population and housing, but mostly housing data. These were tabulated and summarized down to the individual block level. In 1970 we were talking about roughly 1.7 million blocks throughout the country. Blocked areas included census defined urbanized area, plus smaller jurisdictions that contracted for block statistics.

To meet the needs we have heard strongly expressed for redistricting, particularly from State legislators; we recognized early in the planning process for the 1980 census that we needed to collapse or "telescope" these two sets of tabulations. So that within 1 year after April 1, 1980, Census Day, we could produce total population counts for every legal entity that the Census Bureau recognizes in taking the census—counties, cities, towns, townships, etc., plus all statistical areas that we recognize in the census taking, namely census tracts, enumeration districts, block groups, and individual city blocks. For instance, the State of New York is required by its constitution to insure that the minimum difference between any two assembly districts in the State be no greater than the minimum population of the smallest block on the common boundary between the two districts. In 1970, this required a lot of hectic, and I might say emotional, work on both the part of the Joint Legislative Reapportionment Committee in Albany and in the Census Bureau to put together those figures in a time frame that would enable the people in Albany to perform legislative redistricting. What we are doing and have experimented with is taking the old first count and the old third count from 1970 and producing them simultaneously. As Dave Kaplan categorized it, rather than building the church from the steeple down as we did in 1970, the steeple representing the State and the block the foundation, we are going to build it the right way, starting with the foundation and topping it off with the steeple.

We are hopeful that by the end of this calendar year, we will have produced from the Richmond Dress Rehearsal Census, a test computer tape which you will be able to get a test tape which we are now calling Summary Tape File 1. Summary Tape File · 1 will contain approximately 300 cells of information. Again, gleaned from the questions asked of every household—the 100 percent questions.

We are under obligation under a new Federal law, Public Law 94-171, that was inacted in late 1975, that requires us to provide these very detailed population counts, detailed in terms of small area precision, to the Governor and the legislative bodies of each of the 50 States by April 1, 1981, again 12 months after Census Day. This is collapsing a lot of work into a time period that the Census Bureau has never done since we started the block statistics in 1940. We are determined to apply every feasible effort to carry out this mandate.

Another of the operational changes that we are making that will facilitate and aid small-area data analysis in 1980 is an expansion of the block statistics program. As I said a moment ago, in 1970 we tabulated block statistics for urbanized areas, plus contract areas that paid us, and that amounted to about 1.7 million blocks. For 1980, we have expanded that program. Recognizing the needs of small area users and especially those related to redistricting, we will be producing block statistics data for 2.5 to 3 million blocks. Holding aside the contract block statistics program which was again offered in 1980, there has been a general expansion of the block program over 1970 to include all incorporated places of 10,000 or more population that are outside the census defined urbanized areas.

In terms of the contract block statistics program, between 250 and 300 areas have entered into agreements to get data for their jurisdictions, including five entire States—New York, Virginia, Rhode Island, Georgia, and Mississippi; these States will have block statistics from border to border in 1980. Based on their experience with the courts in trying to perform redistricting after 1970, the legislatures in those five States recognized that they needed very precise figures in terms of population counts for carrying out that function. So those five State legislatures enacted legislation and allocated funds to enter the contract program to get block data for their entire State.

The third major change for 1980, which will have a very profound impact on small area data usage, will be an increase in the size of the sample we are using for distributing what we call the long questionnaire form. In 1980, the short form will include 7 population questions for each member of the household and 12 items concerning the characteristics of the housing unit itself for a total of 19 questions. Nationwide, 80 percent of the households will be asked to answer only a limited number of items and 20 percent will be given a more extensive set of questions covering both the 100 percent and additional items on population and housing.

I said *nationwide* very deliberately, because in 1972 we had a program inacted by the Congress called General Federal Revenue Sharing or the State and Local Fiscal Assistance Act of 1972. One of the elements in the allocation formula for distributing the billions of dollars under that program is per capita income as collected in the census. In 1980, since we collect income only on the long form, we are going to increase the sample size in the less populated governmental jurisdictions to get a more precise statistical picture of what the per capita income levels are in those places under 2,500. When I say place, I mean county, village, town, or similar type of general purpose local government.

One in two households in these small communities will get the long form (50 percent sample). Elsewhere throughout the country, only a one in six sample will get the long form questionnaire. This means not only will we have more precise income figures for these smaller jurisdictions, but also all the other information on occupation, journey to work, industry, migration, and so forth, that are collected on the long form questionnaires. All of those data will be vastly improved in terms of statistical accuracy. They still won't be perfect for some of the very tiny governments. As we have done in the past, we will publish tables for the estimated standard errors, cautioning the users to use that information wisely as they apply these data.

Let me say a few words about special small area data that we are going to be preparing. Under Public Law 94-171, we will be producing data for all or for part of election precincts in about 40 States. Let me caution you when I say data. Some of you will be thinking income, occupation, etc., but in terms of the requirements of the law, we are to produce only the total population count; however, we have gone a bit beyond that requirement. In our meetings with the State legislatures in the first half of 1976, it became clearly evident that we needed to provide counts for major race groups and persons of Spanish origin. These data will be included so the States can carry out the clearance function with the Department of Justice, Civil Rights Division, in terms of new election districts boundaries.

Yesterday I was talking with the National Center for Education Statistics (NCES), and we have underway a planned project to produce data for the approximately 16,000 elementary and secondary school districts in the nation. We have not settled the final specifications on the data items that will be available. They will focus mainly on legally mandated data items that are necessary for allocating funds under Title 1, which is essentially the number of school age children in families with low income. It may go beyond that since NCES has certain other analytical and administrative data needs.

The next major program that represents new small area data will be the production of data for officially recognized neighborhoods. This grows out of legislation which was originally introduced by congressperson Patricia Schroeder of Colorado, who at the time was the chairperson of our oversight committee in the House. We have not issued final criteria, but what is essentially involved in this neighborhood program is the production of the same types of data, both 100 percent and sample, that we tabulate and disseminate for census tracts, very similar to the census tract bulletins of previous censuses. These data would be tabulated and produced in the same type of table image format for each officially recognized neighborhood that meets certain guidelines. Those guidelines will require that neighborhoods not overlap since statistically that doesn't make any sense. Second, that they be officially recognized. Third, that they have some form or mechanism for representation, meaning the people or residents of those neighborhoods can make their views known to the city government on issues that might have an impact on their neighborhoods. Those are the three principal criteria we are now studying.

Finally, we have a program for the production of data for traffic analysis zones. We are discussing with the Department of Transportation about the feasibility and cost of repeating what we did in 1970. The program will be on a voluntary basis with metropolitan areas, and the Census Bureau will produce data for traffic zones that the local agencies define in terms of census geography essentially using the same approach that will be taken with election precincts and school districts. The local officials have to define in terms of whole census areas (not splitting blocks) their specialized areas for which we would then produce data.

In terms of the printed reports, (printed reports are still the major census documents that are used by most users throughout the country), we will be producing detailed statistics for Alaskan native villages, for American Indian reservations, for towns in New York and Wisconsin, townships in Michigan, New Jersey, and Pennsylvania and of course, as we have done in the past, towns in New England.

Lastly, let me mention a new series of printed reports called *Summary Characteristics for Governmental Units,* which will be available from the 1980 Census. This essentially represents a set of summary characteristics, both 100 percent and sample population and housing data for each of the general purpose local governments that are eligible for revenue sharing funds and similar programs. We not only plan to produce this series of reports, but we plan to mail a copy to the highest elected official of each of these jurisdictions.

The Census Bureau's Marketing Program: Helping Users Access and Use Census Products

Michael G. Garland
Bureau of the Census

Marketing is defined by the American Marketing Association as "the performance of business activities that direct the flow of goods and services from producer to consumer or user." While the marketing function is readily associated with private sector businesses (as suggested in the preceding definition), it is seldom thought of as a public sector activity. However, governmental agencies whose mission includes providing goods or services to the public must also engage in some form of marketing to successfully fulfill their charge. This paper describes the marketing program of one of these agencies—the Bureau of the Census.

The Congress has directed the Census Bureau, through title 13 of the U.S. Code, to collect and publish statistics on a variety of subjects considered to be of value not only to the Congress and Federal agencies, but also to State and local governments, private enterprises, academic institutions, nonprofit research and community service organizations, trade associations, and neighborhood groups. The intention is to provide these statistics as a means to support sound planning and decisionmaking within our society. However, simply publishing a large number of statistical reports (as many as 4,000 titles a year) on a variety of topics does not achieve this objective.

In order to be faithful to the intent that the data are collected, tabulated, and published to be *used*, rather than just *available*, the Census Bureau seeks to promote increased and improved use of its data by including the marketing function as an integral part of its ongoing activities. It is important to note that increased and improved use of the data is thought of as not only facilitating sound planning and decisionmaking in the Nation's public and private sector organizations but also as achieving a "fair return" on the public investment in these data.

The Bureau's marketing program consists of a core group of activities similar to those associated with private sector marketing efforts—research, promotion, distribution, education, and support. The remainder of this paper describes how the Bureau, as a public agency, carries out each of these five marketing activities in order to achieve the objective of helping users know about, acquire, understand, and use Census Bureau products and services.

Marketing Research

Even though the term "marketing research" may be an uncommon one at the Census Bureau, the function is carried out in a variety of ways to learn of users' needs. For example, 1980 Census planners participated in 73 "public hearings" around the country, attended by more than 6,000 persons, to obtain suggestions about data content, products, and services. They continue to draw upon the Federal Agency Council for Demographic Censuses and the more than 10,000 subscribers of *1980 Census Update* for recommendations on various aspects of the census. Planning for the economic censuses included soliciting suggestions from government agencies, individual firms, labor unions, and literally hundreds of trade associations. Marketing research is also carried out through surveys such as the 1976 survey of summary tape processing centers, the 1976 pilot survey of business users of statistics, and user surveys of particular products such as the *Statistical Abstract* and the *General Report on Industrial Organizations* from the 1972 Enterprise Statistics program. And, of course, the nine Census advisory committees are an integral part of the Bureau's marketing research program. Through these and other efforts, we seek to determine what data are needed, how they should be presented, how our products can be most easily accessed, and what information and assistance is needed to support their use.

Product Promotion

"Oh! I didn't know the Census Bureau collected data on..." is a phrase heard over and over by Bureau staff members. The promotion function of the marketing program seeks to introduce potential users to the Bureau's entire product line and to inform experienced users of new products that may be of interest. Even though paid advertising is not used, a variety of approaches are taken to increase user awareness of the various products and services available.

The exhibit program, for example, annually features exhibits at more than 50 trade and professional conventions, providing opportunities to distribute information to more than 100,000 current or potential data users. Announcements of selected new products are mailed to such organizations as State and regional planning and economic development agencies, universities (e.g., business and economic research, demography, and urban studies centers), trade, and professional associations. Product announcements are also mailed to a list of more than 7,000 names maintained by the Bureau. Products and services are described in the Bureau's monthly newsletter, *Data User News*, and the *Bureau of the Census Catalog.* Attention is also called to the Bureau's statistics through the public information program. Press releases on most major reports are regularly sent to a large number of news media organizations as well as trade and professional journals and other specialized media. Finally, brochures, pamphlets, and other descriptive materials are prepared to inform data users about Bureau programs and products.

Product Distribution

Census Bureau products are distributed through a variety of channels; the U.S. Government Printing Office, the Government's "retail store" for publications, and the local library. There are more than 1,000 Federal Depository Libraries around the country which receive copies of all or selected Bureau publications, supplemented by more than 100 Census Bureau Depository Libraries. Additionally, most other major public and academic libraries maintain copies of Census reports, at least for their local area. City planning departments, chambers of commerce, and regional planning and economic development organizations also frequently have copies of census reports for public use.

Data are sold by the Census Bureau on microfilm, microfiche, and computer tape. Census Bureau public use tapes are also available from more than 75 public and private organizations identified with the Census Bureau's Summary Tape Processing program and other organizations. As a major component of the new State Data Center program, discussed more fully below, the Census Bureau will provide tapes, printed reports, maps, and other products free of charge to the State centers to be used in their data dissemination programs. The centers will provide tape copies and printouts to users, maintain reference libraries, answer user inquiries, and provide analytical assistance involving Census Bureau produced statistics.

User Education and Training

A comprehensive marketing program for a line of products as varied and complex (in a technical sense) as the Bureau's—necessarily includes an education and training component. The Census Bureau meets this need through a broad educational program consisting of a variety of conferences, seminars, workshops, and courses. Survey courses targeted towards State and local government personnel and librarians are held several times a year. Workshops are offered on such topics as accessing and using the Bureau's economic statistics, using machine-readable data products, making population estimates, and using census data to meet Federal legislative and administrative requirements. As an integral part of the economic, demographic, and agriculture censuses, staff members participate in conferences around the country to familiarize data users with the availability and use of the products and services coming from these programs.

An important related activity is the college curriculum support project. Under the auspices of this project, materials are prepared for use in college level courses in sociology, geography, business, urban planning, library science, and other disciplines to systematically introduce students to census products. Over 1,000 instructors are using materials developed to date.

Product Support

Users of Census Bureau products frequently have need for detailed background information on particular products or for explanations of how to use them. A considerable amount of supporting information is provided in the educational activities described above. Additional product support is provided through such activities as inquiry and consultative services, preparation of reference materials, and user aids. More than 150,000 copies of the *telephone contact list* have been distributed over the past 5 years, providing an important linkage between data users and more than 150 subject specialists within the Bureau. To help users understand and use Census Bureau products, a variety of guides, indexes, data finders, procedural reports, and other aids have been prepared. Examples include the *1970 Census Users' Guide, Mini-Guide to the 1977 Economic Censuses, Index to 1970 Census Summary Tapes,* and *Directory of Federal Statistics for Local Areas.*

A different type of product support is provided by the Bureau through the preparation and distribution of computer software to facilitate use of both the data and geographic reference files distributed by the Bureau. Programs available include such capabilities as data aggregation and display, computer mapping, geographic coding, and area and point calculation.

Decentralized Marketing Activities

Data users are found in the four corners of the country and everywhere in between. As the Bureau's marketing program grew, it became quite apparent that users could not be served effectively from Washington alone. In fact, the Federal and Census Depository Library programs were established in the mid-1800's and 1940's, respectively, in recognition of this problem. To increase and improve the services available to data users, the Census Bureau has initiated two new programs which place key marketing activities closer to the user.

1. The Regional Data User Services program consists of user services specialists located in each of the Bureau's 12 regional offices. These specialists answer inquiries about census publications and other Bureau products, assist users in the access to and use of census data needed for specific applications, and make presentations to groups interested in the statistical programs and products of the Bureau.

2. The State Data Center program is a cooperative data dissemination and user services program involving the Census Bureau and participating States. Through the combined efforts of State agencies (e.g., planning, community affairs, or administration) and universities, the States will provide such services to data users as inquiry handling, user training, orientation, consultation, library facilities, tape processing, and general analytical support for data use. The Census Bureau will assist the States in developing the capacity for delivering these services by providing material resources such as printed reports, computer tapes, software, and maps; on-site training and technical assistance; and program coordination. When fully implemented it is anticipated that this program will contribute to improved planning and decisionmaking in State and local government agencies, as well as other public and private organizations, as a result of enhanced opportunities and abilities to use statistical data.

The Future

The 1977 Economic Census, 1978 Census of Agriculture, and 1980 Decennial Census present a tremendous challenge to the Bureau's marketing program. Planning is well underway and programs are being developed to be as responsive as possible to the mandate that planners and decisionmakers in the public and private sectors be informed of the availability of the Bureau's products and that access to and use of these products be convenient, timely, and cost-effective. Major efforts in the future, in addition to continually improving all of the activities mentioned in this paper, will be directed towards increasing the awareness and use of Bureau products in the private sector and providing assistance to users in the application of Bureau data to specific research, planning, and decisionmaking problems.

Comments: Small-Area Data From the 1980 Census

Pearl M. Kamer
Long Island Regional Planning Board, New York

As a user of small area census data, I'm delighted that the Census Bureau has decided to expand its program of small-area statistics. The expanded block statistics program, increased sampling of households in smaller jurisdictions, improved population counts for school districts, and the preparation of neighborhood profiles will be particularly useful for local planning purposes.

The need for small-area statistics has increased greatly since the 1970 census. Federal revenue sharing funds are allocated to general purpose local governments on the basis of census small area data. Many federal special purpose grants are now targeted to specific neighborhoods and require detailed information about social, housing, and economic conditions in those neighborhoods. These trends reflect a new awareness that a broad-brush approach to current urban problems is inadequate and what is needed are local solutions by local decisionmakers on a neighborhood scale.

As an economist employed by a regional planning agency which is charged with neighborhood as well as area-wide planning, I'd like to outline some of the ways in which we plan to use small area data from the 1980 census.

We are currently assisting county police and health departments to construct demographic, economic, housing profiles of their respective police precincts, and health districts. By use of DIME coding, we will be able to splice local crime records and health statistics to forthcoming small area census data. This will enable police planners to relate the incidence of crime to income levels, housing conditions, and unemployment rates in given police precincts. Health planners will be able to relate the incidence of given diseases to the socioeconomic characteristics of given neighborhoods.

As part of our efforts to attract industry to Long Island and to maximize local job growth, we plan to construct small area profiles of the resident labor force including their age composition and occupational skills. A similar manual was prepared following the 1970 census and received a favorable response from the business community. The new manual will be prepared for even smaller geographic areas. We anticipate that it will assist incoming firms to select a site within commuting range of their target labor force and that it will help existing firms to pinpoint the location of a labor force which is occupationally suited to their needs.

We plan to utilize the new traffic zone data to develop information on the worktrip patterns of Long Island residents. Following the 1970 census, we analyzed commuter patterns between Nassau-Suffolk and New York City and pinpointed worktrip origins to major employment centers within the Nassau-Suffolk SMSA. These studies revealed that most Nassau-Suffolk residents traveled relatively short distances to their jobs and that virtually all of them drove to work. The high incidence of auto use for worktrips imposes an economic hardship on those who do not drive—the poor, the young, and the elderly—and could cause labor market disruptions if energy constraints become critical. The new traffic zone data will be used to update commuter profiles on Long Island and will provide the basis for instituting bus, minibus, and jitney service, for redesigning existing bus routes, and for facilitating carpooling.

My agency is critically in need of accurate information about the population of local school districts. Long Island's population had stabilized and a number of school closings have already occurred. In order to devise and implement a realistic plan for school consolidations, we need information about the enrollment population of each school district.

These are but a few of the uses to which we plan to put small area data from the 1980 census. We are looking forward to working with Census Bureau personnel to tailor their product to our planning needs.

An Evaluation of the Census Bureau Marketing Program From a Private Sector Point of View

Edward J. Spar
Marketing Statistics, Inc.

INTRODUCTION

Compared to my first experience which was the 1960 census the present attitude and service setup is highly commendable and refreshing. In the past 10 years, especially, the Bureau has changed from a forbidding fortress to an easily contacted and helpful organization. Therefore, any criticisms of the efforts which have been mentioned should be considered in that light.

POSITIVE COMMENTS

After reading these papers, I decided to conduct my own market research with the New York regional office and found them very knowledgeable and cooperative. They in turn wanted to know what a Summary Tape Processing Center (STPC) in the field was doing and visited our offices. They were concerned that we were doing what we said we did. They have on their part been very diligent about census and other tape referrals. When asked about the September 1979 Economic Census conference, they had to do research to find out when, and did so. More on this later. However, they did give some honest answers to my questions.

The training services are excellent. We have sent one programmer to the computer software and machine readable data products sessions and he has reported to have received good training. In fact, someone else is at Suitland for the latest session.

We were gratified that suggestions we made for the 1977 Census of Retail Trade were taken into consideration and had something to contribute and the Bureau agreed. We hope this two way dialogue which in the past was an impossibility will continue. The Census Bureau has gone out of their way to help us with data and scheduling. They have a strong desire to learn what our problems are.

We notice that new publications are promoted better lately. Through Data Users News, Federal Statistical Users Conference (FSUC), *American Demographics*, and others, we know about new reports rather quickly. Although I still find the Bureau's Catalogue difficult to use.

I have been very impressed with the desire, on the Bureau's part to continually add people to their mailing lists and they also added me. This positive marketing approach has helped my organization greatly in receiving timely data.

CRITICISMS

Although the retail trade group have become more responsive, we believe they still have a long way to go with regards to machine readable data products. They have made great strides this year in starting a prepublication tape program. I hope they will continue this and expand it. With proper promotion, the retail trade data would be a popular item in machine readable form. Also, the Bureau might think about courses or manuals on how to use these data.

The Bureau does not seem to be responsive to suggestions with regard to Summary Tape Files. I submitted suggestions for Summary Tape Files 1 and 2 based upon a request for comments from the Bureau. I have no idea if they were used, considered, or merely filed. Some response to suggestions should be a standard operating procedure.

I am very concerned over what I consider to be a misuse of the Data Users News. The Readers Exchange does state that there is no endorsement of outside data sources being made. Still people use it as an advertising vehicle and sometimes it is misused. I would suggest limiting Data Users News to Census Bureau business.

The New York regional office needs more publicity. As mentioned before, they are dedicated people but I believe very few private sector people know they exist. Further, there are only two people in the New York Regional office. I hope the 1980 Census will bring more people into the office.

Finally, on the local 1977 Economic Census meetings, a small item showed up in *American Demographic* mentioning the New York regional office. The New York meeting is 5 weeks away, and I question if anyone in New York City knows about it. Why is there such a lack of interest on the part of the Bureau to promote this important meeting? In general, why does the Bureau not promote this data base more heavily?

☆ U. S. GOVERNMENT PRINTING OFFICE : 1980 311-046/1163

CENSUS TRACT PAPERS

Series GE-40

No. 1. Papers Presented at the Census Tract Conference, September 10, 1965, Philadelphia, Pa.
No. 2. Needs and Plans for the 1970 Censuses (Census Tract Conference of August 15, 1966, at Los Angeles, Calif.)
No. 3. Some Uses of Census Tracts in Private Business, by Wilbur McCann (1967)
No. 4. Papers Presented at the Conference on Small-Area Statistics, American Statistical Association, Washington, D.C., December 27, 1967, and Related Papers
No. 5. Papers Presented at the Conference on Small-Area Statistics, American Statistical Association, Pittsburgh, Pa., August 23, 1968, and Related Papers
No. 6. Final 1970 Census Plans and Four Programing Systems for Computerized Data Retrieval and Manipulation (Conference on Small-Area Statistics, American Statistical Association, August 21, 1969, New York, N.Y.)
No. 7. New Uses of Census Resources, The Southern California Census Use Study and a Related Paper (Conference on Small-Area Statistics, American Statistical Association, December 29, 1970, Detroit, Mich.)
No. 8. Small-Area Statistics: Strengthening Their Role in Federal Government and Their Use in Criminal Justice Programs (Papers Presented at the Conference on Small-Area Statistics, American Statistical Association, August 23, 1971, Ft. Collins, Colo.)
No. 9. Social Indicators for Small Areas (Papers Presented at the Conference on Small-Area Statistics, American Statistical Association, August 14, 1972, Montreal, Canada)
No. 10. Statistical Methodology of Revenue Sharing and Related Estimate Studies (Papers Presented at the Conference on Small-Area Statistics, American Statistical Association, December 27, 1973, New York, N.Y.)

Copies of early editions are not longer in print but may be available in Federal Depository Libraries. Papers presented at conferences held from 1958 through 1964 were originally published by the Bureau of the Census in its Working Paper Series.

SMALL-AREA STATISTICS PAPERS

Series GE-41

No. 1. Intercensal Estimates for Small Areas and Public Data Files for Research (Papers Presneted at the Conference on Small-Area Statistics, American Statistical Association, August 26, 1974, St. Louis, Mo.)
No. 2. Business Uses of Small-Area Statistics and Education's Needs and Methods for Estimating Low-Income Population (Papers Presented at the Conference on Small-Area Statistics, American Statistical Association, August 25, 1975, Atlanta, Ga.)
No. 3. Statistical Issues in Allocating Federal Funds and Estimation of Local Government Finances (Papers Presented at the Conference on Small-Area Statistics, American Statistical Association, August 25-26, 1976, Boston, Mass.)
No. 4. Interrelationship Among Estimates, Surveys, and Forecasts Produced by Federal Agencies (Papers Presented at the Conference on Small-Area Statistics, American Statistical Association, August 17, 1977, Chicago, Ill.)
No. 5. Methodology and Use of Small-Area Statistics in Decisionmaking and 1977 Economic Censuses and Their Use in Private and Public Sectors (Papers Presented at the Conference on Small-Area Statistics, American Statistical Association, August 15, 1978, San Diego, Calif.)
No. 6. An Evaluation of Small-Area Data Forecasting Models and 1980 Census—Small-Area Statistics Program (Papers Presented at the Conference on Small-Area Statistics, American Statistical Association, August 14, 1979, Washington, D.C.)

For Information on any of these reports, write to:

Subscriber Servcies Section (Publications)
Bureau of the Census
Washington, D.C. 20233

Superintendent of Documents
U.S. Government Printing Office
Washington, D.C. 20402

Official Business
Penalty for Private Use, $300

Postage and Fees Paid
U.S. Department
of Commerce

U.S.MAIL

COM-202

Special Fourth-Class
Rate—Book

Series GE 41, No. 6 An Evaluation of Small-Area Data Forecasting Models and 1980 Census—Small-Area Statistics Program

UNIVERSITY OF ILLINOIS-URBANA
3 0112 084237707

Small-Area Statistics